社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# 前後文脈を考慮したTweetの現地性判断

鬼塚友里絵†　　嶋田　和孝†

† 九州工業大学 情報工学部 知能情報工学科
〒 820–8502 福岡県飯塚市川津 680–4
E-mail: †shimada@pluto.ai.kyutech.ac.jp

**あらまし**　Web 上には多くの観光情報が存在し，その情報を分析することは観光情報学における重要なタスクの一つである．本論文では，Twitter に投稿された各 tweet の現地性を判断する手法について提案する．ここで現地性とは，その tweet が実際に現地で現地のことを言及しているかと定義する．tweet の現地性を正確に判断できれば，観光地に対する評判の分析や移動経路の推定などが可能になる．提案手法では，ルールによるフィルタリングと機械学習によって現地性を判断する．2 つの手法を統合的に利用することで，精度が向上することを示す．また，前後の Tweet に関する情報を利用することの有効性についても検証する．
**キーワード**　Twitter，前後情報，現地性判断

# On-site likelihood identification of tweets with context information

Yurie ONITSUKA† and Kazutaka SHIMADA†

† Department of Artificial Intelligence, Kyushu Institute of Technology
680–4 Kawazu Iizuka Fukuoka, 820–8502 Japan
E-mail: †shimada@pluto.ai.kyutech.ac.jp

**Abstract**　The Web contains much information for the tourism, such as impressions and sentiments about sightseeing areas. Analyzing the information is a significant task for tourism informatics. A useful target resource for the analysis is information on Twitter. However, all tweets with keywords, which are related to facilities and events for tourism, might be not tourism information. In this paper, we propose a method for estimating on-site likelihood of tweets. The task is to identify whether each tweet has high on-site likelihood. We introduce a filtering process and a machine learning technique for the task. In addition, we apply previous and next tweets for the identification task, as context information. Experimental results show the effectiveness of the combination method and context information.
**Key words**　Twitter, Context, On-site likelihood

## 1. ま え が き

観光は成長産業として大きな期待が寄せられている分野であり，国家プロジェクトとして観光振興が実施されている．効果的な観光振興を行なうためには，観光者がどのような観光を行なっているのか，すなわち，観光の実態調査を行ない，収集された情報に基づいて新たな観光戦略や施策を構築する必要がある．一方，様々な観光地を訪れる際に，Web 上のコンテンツを用いて自らの行動を日記として発信するユーザが増加している．このような環境下で，Web 上に存在する観光情報の分析は重要な役割を担いつつある [7], [11]．しかし，Web 上には膨大な量の情報があり，それらすべてを人間が分析するのは現実的でない．そこで，Web コンテンツを対象とした観光情報マイニングに関する研究が活発になってきている．郡ら [4] は，地名と格助詞に着目し，ブログを対象に行動経路の抽出と視覚的提示を行なっている．ブログには，「フォートラベル(注1)」や「旅行・観光ブログ村(注2)」など観光に特化したコンテンツも存在する．このようなコンテンツには良質な観光情報が記載されているが，利用しているユーザは多くはない．

この問題に対応するため，本研究ではマイクロブログの 1 つである Twitter(注3) に着目する．Twitter は，近年ユーザ数が激増しており，新たな情報源として注目されている [10]．Twitter

(注1)：http://4travel.jp/
(注2)：http://travel.blogmura.com/
(注3)：https://twitter.com/

の利点としては，多数のユーザが存在することに加えて，高いリアルタイム性が挙げられる．これを観光情報という視点から見れば，観光地で携帯端末から気軽に投稿された感想を収集可能だということになる．また，手軽に投稿できることからユーザは感じたままに投稿を行うため，従来のブログでは語ることを躊躇するような飾らない意見を収集できるという利点もある．この高いリアルタイム性により，Twitter は観光地の「今」を知るには最適である．一方で，tweet に観光地名が含まれるからといって，必ずしも抽出された情報が分析に必要なものとは限らない．例えば,「今度 ⟨⟨ 観光地名 ⟩⟩ に行きたいな.」という投稿は，投稿者の観光地での「今」の状態を表すものではない．投稿者が観光をしていることを前提としても，その投稿者の他の tweet には，雑談などの多くのノイズが含まれる可能性がある．したがって，各 tweet が実際に現地での体験を表現しているかを判定する必要がある．

本研究では，Twitter から収集した観光情報を対象として現地性判断を行なう．ここで現地性判断とは，ある投稿が投稿者がその場にいて体験したことに関するものかどうかを識別することである．現地性が判断できれば，tweet から観光情報として適切なもののみが収集可能になり，適切な分析が行えるようになる．また，現地性のある tweet を分析すれば，その投稿者の観光地での行動経路や複数の投稿者の情報を統合して典型的なもしくは意外な行動パターンの抽出などが可能になる．このように現地性判断はさまざまな分析の前処理として極めて重要である．我々は先行研究において，ルールベース手法と機械学習を組み合わせた現地性判断の手法を提案し，その有効性を示している [15]．一方でこの先行研究では，観光地名を含む投稿を無作為に収集したものを対象データとし，単純に投稿に現地性があるかどうかのみに着目していた．すなわち，tweet の前後関係やそもそも本当に観光しているのかという前提を考慮していなかった．本稿では，対象を少なくとも特定の場所で観光しているユーザとし，そのユーザから 24 時間中につぶやかれた各 tweet が観光地での体験などを表すつぶやきであるかを判定するタスクとして現地性判断を行う．各ユーザの時系列情報が利用できるため，対象となる tweet の前後の情報を利用することが可能になる．このような前後の情報と先行研究のルールの改良などにより，現地性判断の精度向上を目指す．提案手法の流れを 1 に示す．提案手法ではまず二段階のルールを適用し，現地性を判断する．このルールはフィルタリングとして機能するため，高い精度が必須である．このルールの適用では現地性を判断できなかったものについては，機械学習を適用し，最終的な現地性判断を行う．

## 2. 関 連 研 究

前述のように Twitter の利点は，現地性と高いリアルタイム性である．荒牧らによるインフルエンザの予測 [1] や榊らによる地震情報に関する分析 [12] はこの Twitter の利点を有効に利用した研究である．Cheng ら [2] や Eisenstein ら [3] は，tweet を対象としたユーザの位置情報などの推定や分析を行っている．また，乾ら [6] の経験マイニングに関する研究や成田ら [9] による事実性解析の研究も正確には現地性とは異なるが，同じような視点からの研究である．

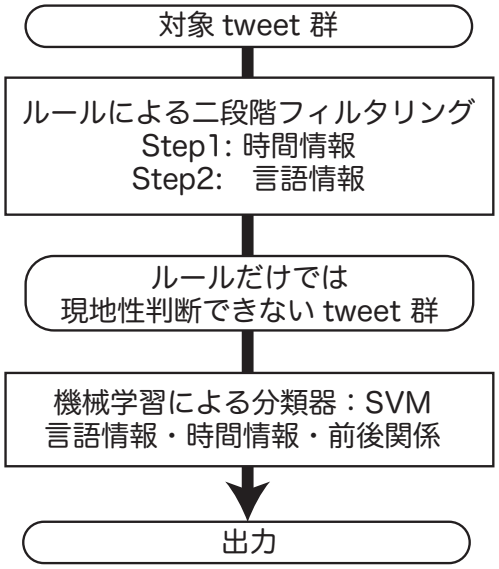


図 1 提案手法の概要.

佐々木ら [13] は tweet を対象として場所参照表現（特定の場所を指し示す表現）を実際の地図などと対応づける研究を行っている．宮部ら [8] が場所依存記録と呼ぶものは本研究での現地性判断と近いものである．彼女らは，つぶやきに含まれる形態素の n-gram を素性として SVM に適用し，この問題を解いたが，一般に全 tweet 中で現地性があるものは，現地性がないものに比べて少ないことが多く，アンバランスな学習データによって適切な学習器が生成されないことが我々の以前の研究 [15] より分かっており，本論文での提案手法もこの問題点に対応している．

## 3. 対象データ

本研究で対象とするデータは，ユーザが観光地にいることを前提とする．したがって，まず，任意の観光地にいることが保証される tweet を収集する必要がある．そこで，Twitter の API を利用して，任意の観光地名を含む tweet を収集する．この時点で，ある tweet が実際にその観光地でつぶやかれているかどうかを人手で判断する．次に，その観光地でつぶやかれた tweet のユーザのタイムライン（前後の tweet）を API を利用して取得する．この取得された各ユーザのタイムライン上の各 tweet に対して現地性の有無をタグ付けし，これを学習・実験の対象データとする．

図 2 にデータ取得の流れを示す本論文ではキーとする観光地名を「清水寺（京都）」とし，データを収集する．API によって獲得されたさまざまなユーザのつぶやきについて実際に「清水寺」にいるユーザのみを特定し，そのユーザのタイムラインを再び API を使って獲得する．図の例では，ユーザ A，B，C の tweet は「清水寺」というキーワードを含んでいるが，それらは願望や実際の観光地でのつぶやきではないことがわかる．一方で，ユーザ D の「夏美と清水寺なうう」という tweet は

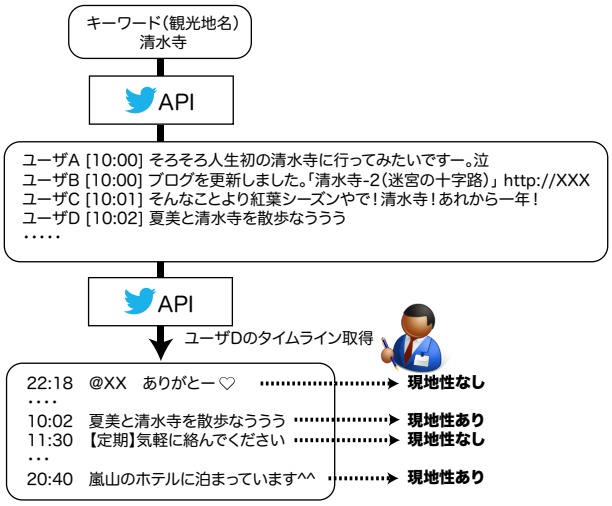


図 2　データ取得の流れ.

現地での行動を表しており，現地性ありと判断される．そこで，ユーザ D のタイムラインを獲得し，このタイムライン中（前後 12 時間）に含まれる各 tweet に現地性についてのタグ付けを行って実験に利用する．

## 4. 提 案 手 法

本節では，先行研究 [15] をベースに，それを改良した提案手法について述べる．現地性判断の手法は (1) ルールによる処理と (2) 機械学習による分類器の 2 つに大別される．ルールではその適用方法を，機械学習では素性の追加によって先行研究の手法を拡張する．

### 4.1 ルールによる処理

ここでは，ルール処理を用いたフィルタリングについて述べる．収集した Twitter データには，明らかに現地性のないものが多く見られる．これらを除外するため，ルールベース処理を用いて識別を行なう．先行研究では，すべてのルールを一段階で適用していたのに対し，本研究では，時間帯という制約は観光というタスク設定上，他のルールよりも強力に働くであろうという過程から，まず時間帯に関するルールによってフィルタリングを行なった後に，tweet 中の言語情報や Twitter 特有の情報を用いる二段階のフィルタリングを行なう．

ルール処理は以下の流れで適用される．

**ステップ 1**： 時間帯によるフィルタリング

深夜帯 (0〜4 時とする) の投稿を除外する．観光者の深夜帯の投稿は，雑談や 1 日の観光を振り返る投稿であると考えられる．

**ステップ 2**： 言語情報や Twitter 特有の情報を用いたフィルタリング

このステップ 2 では，削除ルールと保持ルールの 2 つが存在する．これは，単純に削除ルールのみで tweet を削除すると単純には判定できないような現地性のある tweet が多く削除されてしまうためである．多くの現地性のある tweet がルール処理で誤って削除されてしまうことは，現地性判断全体での精度低下に繋がるため，避けなければならない．したがって，削除ルールに適合しても保持ルールにマッチする場合は，削除しないこととする．

- **削除ルール**
  - 言語情報：次のような言語表現を含む場合は削除する．

  予定：「明日」「来週」「先日」など未来を示す語を含む．

  推定：「だろう」「かも」など予想を示す語を含む．

  伝聞：「らしい」「みたい」など経験が伴わない語を含む．

  会話：「@」または「RT」を用いた会話文.
  - 文長：文長が 100 文字以上の tweet は削除する．
  - 名詞数：名詞を 36 語以上含む tweet は削除する．
- **保持ルール**
  - 場所を起点にした行動の発言：「到着」「着いた」「出発」などの語を含む場合，何らかの場所を目的として行動を起こしている可能性があるため削除しない．
  - 現在の行動を示唆する発言：「〜にいる」「〜を見ている」など現在の行動を示す語を含む場合，投稿者の今の行動を説明しているため可能性が高いため削除しない．
  - 「今」を表す曖昧な発言：「〜なう」など今を表す情報を含む場合,「今」の情報を多分に含んだ曖昧な表現でため削除しない(注1)

### 4.2 機械学習による処理

次に，機械学習に基づく現地性判断の枠組みについて述べる．機械学習の位置づけは，ルールでは判定できなかった tweet の現地性判定である．機械学習の手法としては Support Vector Machine [16] を利用する．実装には Weka を用いた [5].

Twitter が短文による投稿であることに加え，投稿者が観光中であることを考慮すると，前後の発話との関わりが強くなると考えられる．そこで，先行研究で用いられた素性に加え，前後の発話から得られる情報も SVM の素性として用いる．具体的な素性は以下の通りであり，素性群は言語的な特徴，時間的な特徴，前後関係による特徴などに分別される．

- **言語的特徴**
  - Bag of Words：ベースラインとして Bag of words を使用する．Twitter は 1 文が短いため, 単語が文全体に与える影響は強くなると考えられる．あらゆる単語の出現頻度を見ることで，現地性の有無を判断する際に特有の単語が出現していないかを測る．
  - 特定の単語：現地性を特に判断しやすい言語が存在する．「到着」や「〜にいます」などの表現はその場にいる可能性を高くする.「なう」も文脈の考慮により，現地性の判断に利用できる．このような単語に着目し，その単語の有無を素性とする．
  - 形容詞の時制：投稿文に含まれる形容詞の時制に着目する．例えば,「楽しい」という語が現在形であれば,「今」その気持ちが起こっていると考える．しかし,「楽しかった」や「楽しそう」などの時制の変化が起こっている場合は，過去の経験や未来への期待を語っていることになる．よって形容詞の時制を素性とすることで，現地性を測る．
  - 動詞数：観光地に居る場合,「行く」や「着く」,「見る」

(注1)：tweet 中の「なう」の意味は文脈によって異なることが多く，この語が含まれている tweet を単純にルールで識別することは難しい場合が多い．

といった動詞が見られるようになるが，逆にそれ以外の動詞の出現回数は低くなる．そこで動詞数が多いものほど現地性がないと考え，素性として利用する．

– 名詞数：名詞の数が多くになる場合，何らかの説明を行なっていることが多い．説明を行なっている投稿は，個人の意見が反映されるものが少なくなる傾向があったため，この数を素性の 1 つとする．

– 文長：現地性に着目する場合，投稿者は必然的に外出していることになる．外出の際に用いられる主な投稿手段は携帯端末となるが，そのような端末を用いる場合，多少の入力の煩わしさが存在する．そのため，外出先からの投稿は文長が短くなると考え，文長そのものを素性とする．

- **時間的特徴**

– 時間情報：投稿時刻を考慮する．観光地には，観光客が訪れやすい時間が存在する．昼間賑わう遊園地や，夜に盛り上がる花火大会など，単語情報との組み合わせによって，現地性の判断を行うことができると考える．

– 時間帯：昼には外出している人が多く，夜は屋内から投稿する人が多い．このような傾向を利用するため，時間情報を 3 つに分類し，素性とする．5〜9 時を朝，10〜23 時を昼，0〜4 時を夜とする．

- **前後の tweet から得られる特徴**

– 前後の投稿の Bag of Words：Twitter は，短文の投稿が多く見られる．そのため，1 つの話題に対して複数の投稿が行われる傾向がある．よって，前後の投稿との関連性があると考えられるため，素性として追加する．

– 前後の投稿との時間間隔：観光中であることを考慮すると，前後の投稿の現地性があり，かつ時間間隔が短い場合，その投稿も現地性があると考えられる．よって，新たに素性として追加する．

– 前後の投稿の現地性：観光中であれば，現地性のある投稿が連投される可能性がある．そのため，前後の投稿の現地性も素性として追加する．

## 5. 実　　験

提案された手法についての実験を行う．3. 節に従い，データを収集し，現地性の有無についてタグ付けを行い，データセットを作成する．データセットには SVM の学習に用いる訓練データと評価のみに利用するテストデータの 2 種類を用意する．訓練データは 3868 件の tweet を含み，そのうち 754 件が現地性あり，3114 件が現地性なしである．テストデータは同様の手法で 3768 件あり，その内訳は現地性ありが 832 件，現地性なしが 2936 件であった．この訓練データおよびテストデータをそれぞれ評価し，提案手法の有効性を検証する．また，テストデータへのルールの適用だけではなく，訓練データへルールを適用することで，分類器の精度がどのように変化するかを確認する．

### 5.1 結果と考察

実験結果は (1) 訓練データへのルール適用の精度，(2) 訓練データでの現地性判断の精度，(3) テストデータへのルール適用の精度，(4) テストデータでの現地性判断の精度の 4 つに分かれる．ここで，(1) と (2) はクローズな実験（すなわちルールのチューニングや SVM の訓練データを交差検定などで評価）を意味し，(3) および (4) はオープンな実験を意味する．(2) および (4) については，SVM の単純な精度とルールの結果を踏まえた現地性判断全体での精度の 2 つがある．

表 1　提案手法のルール処理の結果.

| 現地性ありの再現率 | 現地性なしの削減率 |
|---|---|
| 0.956 (721/754) | 0.429 (1337/3114) |

#### 5.1.1 訓練データへのルール適用の精度

まず，訓練データに対するルール適用に関する実験結果について述べる．評価尺度には，現地性ありのタグが付与されたデータの再現率，現地性なしのタグが付与されたデータの削減率を用いる．再現率は式 1，削減率は式 2 を用いて計算する．

$$\text{再現率} = \frac{\text{現地性のあるデータで正しく識別された数}}{\text{現地性ありのタグが付与されたデータ数}} \quad (1)$$

$$\text{削減率} = \frac{\text{現地性のないデータで正しく識別された数}}{\text{現地性なしのタグが付与されたデータ数}} \quad (2)$$

実験結果を表 1 に示す．ルール処理では，33 件の現地性ありの tweet を取りこぼしてしまったが，機械学習の分類において悪影響を及ぼす可能性のある現地性なしの tweet を 1337 件削除することができた．ルール処理は，現地性判断の他に，次の処理へのフィルタリングの意味を持っており，高い再現率は必須である(注2)．実験結果より，提案手法はフィルタリングとして十分機能していると考えられる．

提案手法では先行研究と異なりルールを二段階で適用した．段階適用しない先行研究と同じ条件で実験したところ，提案手法の再現率は先行研究と比べて 1%程度低下したが，削減率については約 25%向上した．これらの結果の現地性判断全体での有効性について次節で議論する．

#### 5.1.2 訓練データでの現地性判断の精度

次に，5. 1. 1 節の処理でフィルタリングを通過した現地性ありのタグが付与された tweet 721 件および現地性なしのタグが付与された tweet 1777 件を含む 2498 件のデータに対して評価する．これらのデータを SVM で 10 分割交差検定により評価した．評価尺度には，現地性ありのタグが付与された tweet に対する適合率と再現率を用いる．適合率は式 3，再現率は式 4 により計算する．

$$\text{適合率} = \frac{\text{正解数}}{\text{現地性ありと識別されたデータ数}} \quad (3)$$

$$\text{再現率} = \frac{\text{正解数}}{\text{現地性ありのタグが付与されたデータ数}} \quad (4)$$

まず，4. 2 節で述べた各素性の有効性について検証する．実験結果を表 2 に示す．表は Bag of words をベースラインとし，それに別の素性を加えた場合の再現率および適合率を表している．例えば，「Bag of words + 特定の単語」は Bag of words

(注2)：すなわち，この処理で失われた 33 件は次の SVM での処理ではそもそも扱えない．

表 2　機械学習による 10 分割交差検定.

| 特徴量 | 適合率 | 再現率 |
|---|---|---|
| Bag of words | 0.621 | 0.460 |
| Bag of words + 特定の単語 | 0.621 | 0.460 |
| Bag of words + 時間情報 | 0.645 | 0.487 |
| Bag of words + 時間帯 | 0.650 | 0.472 |
| Bag of words + 文長 | 0.689 | 0.631 |
| Bag of words + 形容詞の時制 | 0.621 | 0.460 |
| Bag of words + 動詞数 | 0.641 | 0.464 |
| Bag of words + 名詞数 | 0.693 | 0.643 |
| Bag of words + 前後の投稿の Bag of words | 0.597 | 0.485 |
| Bag of words + 前後の投稿との時間間隔 | 0.619 | 0.462 |
| Bag of words + 前後の投稿の現地性 | 0.650 | 0.628 |
| 全素性 | **0.694** | **0.761** |

表 3　現地性判断全体での評価.

| 手法 | 適合率 | 再現率 |
|---|---|---|
| [手法 1] 提案手法 | 0.694 | **0.721** |
| [手法 2] 提案手法（ルール処理なし） | 0.670 | 0.640 |
| [手法 3] 先行研究（二段階・文脈情報なし） | **0.699** | 0.658 |

と特定の単語に関する素性を組み合わせた場合の SVM の精度である．この結果より，名詞数や文長，前後の投稿の現地性などが素性として有効に機能していることが分かる．名詞数や文長が有効に機能するのは，観光中には「今」の状況を簡潔に短く表現する傾向が見られたためだと考えられる．また，ユーザは観光中であるため，自分の観光体験を連続して投稿する傾向がみられた．その結果，前後の投稿の現地性を用いることが精度向上に繋がったものと考えられる．最終的にはすべての素性を利用する場合に適合率・再現率の両面で最も良い結果が得られた．

ここで，表 2 に示した結果は，5.1.1 節から与えられたデータについて SVM で 10 分割交差検定した場合の単純な精度であり，ルール処理で欠落した現地性のある tweet が含まれていないことに注意しなければならない．すなわち，正確に評価するためには 5.1.1 節で欠落した 33 件（現地性ありの約 4%に相当）を考慮しなければならない．そこで，この点を考慮して適合率および再現率を計算した．これはすなわち，ルールと SVM 全体を 1 つの処理として考えた場合の現地性判断の評価である．

実験結果を表 3 に示す．表中の「提案手法」がルールと SVM を組み合わせた手法を表している．「提案手法（ルールベース処理なし）」とは，ルール処理を適用せずに，オリジナルの訓練データ 3868 件を 10 分割交差検定した場合の精度である．「先行研究（二段階・文脈情報なし）」とは先行研究の手法をそのまま適用した場合を意味する(注3)．[手法 1] と [手法 2] を比較した場合，ルール処理を適用する提案手法は，しない場合と比べて再現率が 2.4%，適合率が 8.1%も上昇している．これは，たとえルールの適用で多少の再現率が下がったとしても，明らかなノイズの適切な除去やアンバランスさの問題を完全ではないが

(注3)：すなわち，ルール適用と SVM は併用するが，ルールの多段階化や前後 tweet の情報を利用していない手法.

表 4　テストデータへのルール処理の結果.

| 現地性ありの再現率 | 現地性なしの削減率 |
|---|---|
| 0.942 (784/832) | 0.533 (1567/2936) |

表 5　テストデータに対する現地性判断の結果.

| 手法 | 適合率 | 再現率 |
|---|---|---|
| [手法 1a] 提案手法 | 0.696 | 0.649 (0.600) |
| [手法 2a] 提案手法（ルール処理なし） | 0.614 | 0.626 |

改善したデータに対して SVM を適用する方が全体として有効に機能することを示しており，ルールと SVM を併用することの有効性を示す結果である．提案手法 [手法 1] と先行研究 [手法 3] を比較すると，適合率は誤差程度の差しかないにもかかわらず，再現率は 6%程度の改善がみられた．この実験結果より，ルールの多段階化や前後情報の有効性が確認された．

**5.1.3**　テストデータへのルール適用の精度

5.1.1 節や 5.1.3 節は，クローズドな評価の結果である．したがって，特にルール処理が訓練データにオーバーフィットしている可能性もある．そこで，未知のテストデータに対して，ルール処理を適用した場合の精度を評価する必要がある．

表 4 はテストデータ 3768 件に対して，ルール処理を適用した結果である．評価尺度には，5.1.1 節と同様に現地性ありのタグが付与されたデータの再現率（式 1），現地性なしのタグが付与されたデータの削減率（式 2）を用いる．実験結果から分かるように，訓練データの場合と比べ，現地性のない tweet は多く削減することができたが，一方で現地性ありの tweet の再現率は 1%程度低下した．すなわち，ルールがいくぶん過剰に働いたことを示している．ルールの適用が最終的に現地性判断全体では有効に機能することは前節で示したが，再現率の低下はフィルタリングという特性上好ましくない．より大規模なデータを用いて，ルールの制定について議論する必要がある．

**5.1.4**　テストデータでの現地性判断の精度

最後に，テストデータに対する現地性判断について述べる．学習にはオリジナルの訓練データ 3868 件を利用し，評価には，ルール処理を経て得られた 2153 件（現地性あり 784 件と現地性なし 1369 件）とルール処理を利用しなかった場合の 3768 件（テストデータそのもの）の 2 種類を用いる．したがって，前者は表 2 の提案手法 [手法 1]，後者は表 2 の提案手法（ルール処理無し）[手法 2] を意味する．評価尺度には，5.1.2 節と同様に現地性ありのタグが付与されたデータの適合率（式 3）と再現率（式 4）を用いる．

表 5 に実験結果を示す．表の中で示されている値は，SVM の出力結果としての精度を意味しており，表 2 の全素性の評価に類するものである．一方で，同じく 5.1.2 節で述べたように，ルール処理を適用する場合，現地性判断全体としては，ルール処理で欠落した現地性ありの tweet についても考慮して再現率を計算する必要がある．表中の括弧内の再現率はこの現地性判断全体（ルール処理での誤りを考慮）での評価である．すなわち，この値は表 3 と見比べるべき値となる．結果から分かるよ

表 6　訓練データへのルール適用の効果.

| 手法 | 適合率 | 再現率 |
|---|---|---|
| [手法 1b] ルールあり訓練 | 0.700 | 0.696 (0.650) |

うに，提案手法はルールなしと比べると適合率については大幅に改善したが，現地性判断全体でみた場合の再現率はルールを適用せずに機械学習のみで処理する方が高くなった（0.600 vs. 0.626）．これは前節で述べたルール処理の再現率低下とも密接に絡んでおり，適切なルールの作成が提案手法において最も重要な課題であることを表している．

### 5.2　ルール適用の訓練データへの効果

これまでの手法では，学習された SVM に適用する前処理としてのルール処理であった．一般に，アンバランスな訓練データと比較して，バランスの取れたデータの方が機械学習によって適切な分類器が作成される傾向がある．これはすなわち,「訓練データを作成する」という点についても，ルールの適用が有効に機能する可能性を示している．そこで，オリジナルの学習データについてルール処理を適用した結果（すなわち 5.1.1 節を通過したデータ数 2498 件）を使って SVM を学習し，ルール適用した場合（表 5 の [手法 1a] が処理した対象）の精度を検証する(注4)．

実験の結果を表 6 に示す．数値の意味は表 5 と同じであり，この 2 つの表の値は直接比較することができる．2 つの表の [手法 1a] と [手法 1b] を比較すると，訓練データにもルールを適用し，アンバランスさの解消とノイズ削減を行ったデータに基づいて SVM を学習した [手法 1b] の方が特に再現率の面で有効であることがわかる．さらに現地性判断全体での数値（すなわち括弧内の数値）も表 5 で [手法 1a] を上回っていた [手法 2a] よりも良くなっている．この結果は，ルール処理を学習の段階でも利用することが全体の精度向上に繋がることを示している．

## 6. ま と め

本論文では，Twitter を対象とし，各 tweet に現地性があるかを判断する手法について提案した．現地性判断にはルールによるフィルタリングと SVM による学習の 2 つを組み合わせて利用した．提案手法では，先行研究の前提条件を変更し，(1) ルールの多段階化および (2) 前後文脈の追加，の 2 点について改良した．実験により，ルールの多段階化も SVM への前後文脈の追加も有効に機能することを確認した．さらに，ルールを現地性判断の処理だけではなく，SVM の訓練データ作成にも利用することで，全体の精度が向上することも確認した．

今後の課題としては，まずルールの拡張・改善が挙げられる．ルール処理での失敗（特に抽出漏れ）は現地性判断全体の精度に大きく影響する．汎用性のあるルールを拡充することは重要な課題である．今回は単純に現地性の判断を行ったが，たとえばこの結果をユーザの行動経路抽出などに応用することを考えれば，佐々木ら [13] の研究のように，その場がどこであるのかを正確に推定する必要がある．我々は複数の情報源から最適な観光地を推薦するシステムを構築している [14]．今回の現地性の判定結果をこのような応用システムへ適用していくことも今後の課題である．

(注4)：[手法 1a] が SVM の学習にオリジナルの訓練データ 3868 件を利用している点が本節の実験設定と異なる．

## 文　献


[1] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[2] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–769, 2010.

[3] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287, 2010.

[4] 郡宏志, 服部峻, 手塚太郎, 田島敬史, 田中克己. ログからのビジターの代表的な行動経路とそのコンテキストの抽出. 電子情報通信学会技術研究報告 IEICE-DE, pp. 29–34, 2006.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update: Sigkdd explorations. *ACM SIGKDD Explorations Newsletter*, Vol. 11, No. 1, pp. 10–18, 2009.

[6] 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. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 314–321, 2008.

[7] 伊藤直哉. 観光情報の国際的動向: Fitt 活動を中心に. 人工知能学会誌, Vol. 26, No. 3, pp. 226–233, 2011.

[8] 宮部真衣, 北雄介, 久保圭, 荒牧英治. マイクロブログから場所依存の様相記録を抽出する :“100ninmap”プロジェクトによる街歩きイベントの実施と応用. 言語処理学会 第 20 回年次大会 発表論文集, pp. 420–423, 2014. C2-5.

[9] 成田和弥, 水野淳太, 乾健太郎. 日本語事実性解析課題の経験的分析. 情報処理学会研究報告 第 204 回自然言語処理研究会 2011-NL-204, pp. 1–8, 2011.

[10] 奥村学. マイクロブログマイニングの現在. 電子情報通信学会技術研究報告 IEICE-NLC2011-59, pp. 19–24, 2012.

[11] 斎藤一. Web における観光情報提供と分析. 人工知能学会誌, Vol. 26, No. 3, pp. 234–239, 2011.

[12] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquakeshakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web (WWW2010)*, 2010.

[13] 佐々木彬, 五十嵐祐貴, 渡邉陽太郎, 乾健太郎. 場所参照表現のグラウンディングに向けて. 言語処理学会 第 20 回年次大会 発表論文集, pp. 177–180, 2014. P3-9.

[14] 嶋田和孝, 上原尚, 遠藤勉. 集合知に基づく観光地推薦システムの構築. 観光情報学会「観光と情報」, Vol. 10, , 2014.

[15] Kazutaka Shimada, Shunsuke Inoue, and Tsutomu Endo. On-site likelihood identification of tweets for tourism information analysis. In *Proceedings of 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM 2012)*, 2012.

[16] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1999.