情報処理学会論文誌　データベース　Vol.9　No.2　11–22 (June 2016)

# Twitter投稿文章とプロフィール情報を用いたPOI公式アカウント分類手法

落合 桂一[1,2,a)]　山田 渉[1]　深澤 佑介[1]　菊地 悠[1]　松尾 豊[2]

受付日 2015年12月19日, 採録日 2016年4月8日

**概要**：本研究では，Twitter の投稿文章とプロフィール情報から生成した特徴量に基づき，機械学習により Point-of-Interest（POI）の公式アカウントを判定する方法を提案する．公式アカウントを判定するため，あらかじめ用意した POI データベースを使い POI 名称と Twitter のユーザ名を比較する方法では，1）POI データベースにない POI は抽出されないという課題，2）POI 名称が正式名称のために通称や略称などが使われるユーザ名と一致しないという課題，3）一般ユーザがユーザ名に POI 名称を利用している場合があるという課題がある．そこで，Twitter の投稿内容やプロフィール情報に基づいて POI 公式アカウントを判定する手法を提案する．本研究では，POI 公式アカウント抽出のための特徴量として，従来用いられていた投稿文章や自己紹介文の Bag-of-Words に加え，POI 固有特徴量（場所情報，営業時間，連絡先など），知名度に関する特徴量（フォロワ数やリストに登録されている数など），プロフィール画像の画像特徴量を提案する．実験により POI データベースを利用した場合と比較し，約 3 倍の POI 公式アカウントが抽出可能であることを示した．また，提案した特徴量を利用した場合，従来手法の特徴量を利用した場合と比較し分類性能を表す再現率 0.933，F 値 0.938 で最大になることを示した．

**キーワード**：Twitter，Point-of-Interest，ユーザ属性推定

# POI Official Account Classification Method Using Twitter Posts and Profile Information

Keiichi Ochiai[1,2,a)]　Wataru Yamada[1]　Yusuke Fukazawa[1]　Haruka Kikuchi[1]　Yutaka Matsuo[2]

Received: December 19, 2015, Accepted: April 8, 2016

***Abstract:*** In this paper, we propose a machine learning method for classifying Point-of-Interest (POI) official twitter account using Twitter posts and profile information. There are three issues to classify POI official account by using prepared POI database. 1) POIs which are not in POI database cannot be extracted, 2) POI name which is in POI database doesn't match the name of twitter account which is often abbreviated, 3) even if the user name matches the POI name, in some cases, the account may not be POI official account. Therefore, we classify these accounts by using twitter posts and profile information. In this study, we use POI-related information such as location, business hour, and contact information etc., popularity such as number of followers, listed counts etc., and profile image as features for machine learning in addition to conventional method (Bag-of-Words of tweets and profile text). We evaluated our method with the method which uses POI database. The result showed that proposed method can extract three times as much account as baseline. In addition, the result which we evaluated classification performance showed that the recall and F-measure are higher than that of conventional method.

***Keywords:*** Twitter, Point-of-Interest, User Attribute Estimation

[1] 株式会社 NTT ドコモ
NTT DOCOMO, INC., Yokosuka, Kanagawa 239–8536, Japan
[2] 東京大学
The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[a)] ochiaike@nttdocomo.com

## 1. はじめに

スマートフォンの普及により地域情報を検索するローカル検索を利用するユーザが増加している．ローカル検索で

は，ユーザは施設（Point-of-Interest，以下 POI）の住所や連絡先，営業時間などを検索することができる．Google が行った 2013 年の調査によるとスマートフォン利用者の約 80%がローカル検索を行った後に，問い合わせ，来店，購入や予約といった何らかの行動をとっており，ローカル検索が行動のきっかけとなっている[*1]．一方，Twitter や Facebook などのソーシャルネットワーキングサービスが普及している．Twitter では一般ユーザだけでなく有名人や企業，観光スポットや商業施設が運営するアカウントなど様々なユーザが投稿している．観光スポットや商業施設などの POI の公式アカウント（以下，POI 公式アカウント）では，イベント告知やセール情報などを投稿しており，ローカル検索に比べて即時性の高い情報を得ることができる．そのため，POI 公式アカウントの投稿を位置情報に基づいて検索することができれば有用である．たとえば，Twitter 公式サイトでは有名な動物園・水族館や美術館・博物館の公式アカウントの最新の投稿を表示するページ[*2]を提供している．また，株式会社デジタルガレージが運営するツイナビという情報提供サイトでは，店舗・商業施設の公式アカウントのランキングを提供している[*3]．これらの Web サイトで提供されている施設の公式アカウントの投稿は，リアルタイムな施設情報を取得することができるため，施設情報の 1 つとして見ることができれば有用である．本研究ではこのような利用シーンを想定しており，検索サービスやポータルサービスを提供するサービス提供事業者が本研究の手法を利用することを想定している．そのため，対象となるアカウントが多く存在し，人手ではなく機械的に判定できることが望ましい．

POI 公式アカウントを特定するナイーブな方法として，タウンページ[*4]や Open Street Map[*5]などの POI データベース（以下，POI DB）を使い，POI 名称と Twitter のユーザ名の一致により POI 公式アカウントを判定したり，検索エンジンや Twitter 公式サイトを利用して POI 名称でキーワード検索したりする方法が考えられる．しかしながら，POI DB を利用する方法では，以下 3 点の課題がある．

1) POI DB の POI 登録数に起因するカバレッジ不足に関する課題：POI が POI DB に存在しない場合，当該 POI に関する POI 公式アカウントは抽出されない．
2) POI DB の登録名称とが異なることに起因するカバレッジ不足に関する課題：POI DB に登録されている POI 名称と POI 公式アカウント上で使われている POI 名称が異なる場合，当該 POI に関する公式アカウントは抽出されない．たとえば，「ヴィレッジヴァンガード 池袋サンシャインシティアルタ店」の Twitter 公式アカウントはユーザ名が「V.V. 池袋サンシャインシティアルタ」となっている．
3) POI 公式アカウントの抽出誤りに関する課題：POI DB に登録されている POI 名称と同じ名称が使われている非公式のアカウントがあった場合，当該アカウントが誤って抽出されてしまう．たとえば，Twitter の検索機能で「清水寺」をキーワードにアカウント検索[*6]すると，2015 年 12 月 18 日時点で少なくとも 10 ユーザ以上存在している．また，付録**表 A·1** のスポット名で検索した場合，検索結果の 1 位が公式アカウントであるものは 34 件中 6 件であり，検索結果には一般ユーザも含まれていた．

上記の課題はすべて POI DB を利用するために発生する課題である．そこで本研究では，POI DB を利用せずに Twitter の投稿文章やプロフィール情報（自己紹介文とプロフィール画像）から生成した特徴量に基づき機械学習により判定対象の Twitter アカウントが POI 公式アカウントかどうか 2 値分類する方法を提案する．Twitter の投稿文章や自己紹介文を分析する先行研究として，Twitter ユーザの属性推定の研究がある．従来研究 [1], [2], [3], [4], [5], [6], [7] では，一般ユーザを対象として性別，年齢，職業，興味，居住地，支持政党などを投稿文章や自己紹介文の Bag-of-Words（以下，BoW）[9] から推定する研究が行われていた．本研究では，BoW 特徴量に加え，POI 公式アカウントを抽出するため，POI 固有特徴量（実世界の場所に関する情報，営業時間，連絡先など），知名度に関する特徴量（フォロワ数やリスト登録数など），プロフィール画像の画像特徴量を利用することで BoW のみを利用する従来手法より高性能な手法を提案する．機械学習の特徴量については，2 章で詳しく説明する．

POI 公式アカウントを特定することができれば，POI 公式アカウントの投稿検索だけでなく，POI 公式アカウントの投稿からイベント情報を抽出したり，一般ユーザがフォローやメンションを行っている公式アカウントから一般ユーザの興味や地理的行動範囲などのユーザ属性推定を行ったりするなど様々な応用が考えられる．特に，ユーザ属性推定への応用では，興味と行動範囲の 2 つを同時に推定でき有用であると考えられる．

本研究の貢献は以下のとおりである．

- BoW 特徴量に加え，POI 公式アカウントの特徴を考慮した特徴量として POI 固有特徴量，知名度に関する特徴量，プロフィール画像から生成した特徴量を提案した．
- 実データを利用した実験により，POI DB を利用した

*1 http://adwords-ja.blogspot.jp/2013/07/2011-4.html
*2 https://twitter.com/i/streams/category/686639687419596828
*3 http://twinavi.jp/account/list/%E5%BA%97%E8%88%97%E3%83%BB%E5%95%86%E6%A5%AD%E6%96%BD%E8%A8%AD
*4 http://tpdb.jp/townpage/order
*5 https://www.openstreetmap.org
*6 https://twitter.com/search?f=users&vertical=default&q=%E6%B8%85%E6%B0%B4%E5%AF%BA

名称の一致による手法と比較して機械学習を用いる提案手法のカバー率が高いことを示した．
- 実データを利用した実験により，従来手法より再現率，F 値を向上できることを確認し，追加した特徴量が有効であることを示した．

本稿の構成は以下のとおりである．次章で提案手法について述べ，3 章では提案手法の有効性を確認するために行った評価実験について説明する．4 章で関連研究について述べ，最後に，5 章で本研究のまとめと今後の課題を述べる．

## 2. 提案手法

本章では提案手法について説明する．提案手法の処理の流れを**図 1** に示す．提案手法では特徴量として従来手法 [1], [2], [3], [4], [5], [6], [7] と同様に投稿文章の BoW，自己紹介文の BoW を利用する．加えて，POI 固有の情報を自己紹介文や専用項目に対する正規表現により抽出して生成した POI 固有特徴量，フォロワ数やリストに登録されている数など知名度に関する特徴量，画像特徴量として Bag-of-Visual Words（以下，BoVW）[11] を利用し，教師あり機械学習により POI 公式アカウントを判定する．各特徴量の設計方針は以下のとおりである．

1) POI に関連するアカウントを抽出するため，施設情報と関連する情報（場所，営業時間，問い合わせ先，URL など）を特徴量とする．（POI 固有特徴量）
2) POI に限らず公式アカウントに共通する特徴量を利用する．公式アカウントは一般に知名度が高いと考えられるため，フォロワ数やリストに登録されている数などを利用する．（知名度に関する特徴量）
3) Twitter のプロフィール画像を分析した先行研究 [23] で，ロゴをプロフィール画像に利用しているユーザは公式アカウントである傾向があるという調査結果が出ているため，画像認識の分野で分類に利用される特徴量を利用する．（画像特徴量）

**対象 POI**　ローカル検索のように外出するための情報検索を行う利用シーンを想定しているため，対象とする POI は一般消費者が実世界で利用できる場所とし，観光スポットや商業施設，飲食店などが含まれる．

**POI 公式アカウントの定義**　観光スポットや商業施設を運営する企業，法人や団体が開設した Twitter アカウントとする．一方，同じ企業アカウントでも，一般消費者向けに実店舗を持たない EC サイト，ニュースサイトや地域情報を配信するアカウントは POI 公式アカウントに含めない．

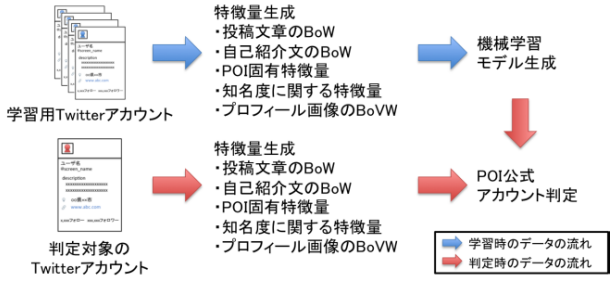



**図 1**　提案手法の処理の流れ
**Fig. 1**　Overview of the proposed method.

### 2.1 従来手法の特徴量

#### 2.1.1 投稿文章の特徴量

既存手法においてユーザの投稿を利用する手法はよく用いられている [1], [2], [4], [5], [7]．本研究でも同様の手法を利用する．本研究では，ユーザごとのツイートを文書の単位とし，ツイートを形態素解析して得られるすべての形態素を特徴量として用いる．BoW のベクトルの要素の値には榊ら [7] と同様に TF-IDF [10] を用いる．

$$tf \cdot idf_{t,d} = tf_{t,d} \times idf_t \tag{1}$$

ここで，$tf_{t,d}$ は文書 $d$ での単語 $t$ の出現頻度，$idf_t$ は $\log(\mathrm{N}/df_t)$ で表される逆文書頻度，N は総文書数，$df_t$ は単語 $t$ が出現する文書数である．

#### 2.1.2 自己紹介文の特徴量

Twitter では自己紹介のために 160 文字以内で文章を記載する自由記述欄がある．自己紹介文の特徴量も投稿文章と同様に形態素解析して得られる形態素を利用する．BoW ベクトルの値には TF-IDF を用いる．本研究では投稿文章と自己紹介文に同じ単語が出現した場合は別の特徴量として扱う．

### 2.2 提案する特徴量

提案手法で利用する従来手法から拡張した特徴量の一覧を**表 1** に示す．以降，各特徴量の詳細と導入した意図を説明する．各特徴量の略称を括弧内に示す．本研究では，Twitter API のレスポンスにおいて “name” の項目に該当するものをユーザ名，“screen_name” の項目に該当するものをスクリーン名と呼ぶ．

#### 2.2.1 POI 固有特徴量

(1) **Location 項目記載有無（Loc-field）**

Twitter にはアカウントの所在地を記載するための location という項目が存在する．POI 公式アカウントでは正確な住所を記載しているユーザがいるため利用する．特徴量は記載有無を 2 値で表す．

(2) **URL 項目記載有無（URL-field）**

POI 公式アカウントではユーザに情報提供することがアカウント開設の目的の 1 つであるため，より情報が豊富な Web ページへのリンクをプロフィールに記載していることが多いと考えられる．そのため，記載有無を 2 値の特徴量として利用する．

(3) **スクリーン名と URL の共通文字列比（Screen-URL-rate）**

スクリーン名は@から始まるユーザ名称であり，英数

表 1　提案手法で追加する特徴量一覧
**Table 1**　List of features of the proposed method.

| 種別 | 特徴量 |
|---|---|
| POI 固有特徴量 | Location 項目記載有無（Loc-field） |
| | URL 項目記載有無（URL-field） |
| | スクリーン名と URL の共通文字列比（Screen-URL-rate） |
| | ユーザ名と自己紹介文の共通文字列比（Name-bio-rate） |
| | ユーザ名の長さ（Name-length） |
| | 自己紹介文中での URL 記載有無（URL-bio） |
| | 自己紹介文中での電話番号記載有無（Tel-bio） |
| | 自己紹介文中でのメールアドレス記載有無（Mail-bio） |
| | 自己紹介文中での営業時間記載有無（Hours-bio） |
| | 自己紹介文中での郵便番号記載有無（Postal-bio） |
| | 自己紹介文中での公式記載有無（Official-bio） |
| 知名度に関する特徴量 | フォロワ数（Followers） |
| | フレンド数（Friends） |
| | 比率（フォロワ数/フレンド数）（FF-rate） |
| | 被登録リスト数（Listed-count） |
| | 認証バッジの有無（Verified） |
| 画像特徴量 | プロフィール画像 Bag-of-Visual Words |

字と記号を利用できる．POI 公式アカウントでは POI の Web ページの URL とスクリーン名が共通していることが多く，その他のユーザとの違いが出やすいと考え利用する．類似度の計算には最長共通部分列比（Longest Common Subsequence Ratio，LCSR）[22] を用いる．共通部分列とは，2 つの系列において，連続不連続を問わず同じ要素が同じ順番で出現した部分列のことである．共通部分列として取り得るもののうち最も長いものを最長共通部分列（Longest Common Subsequence，LCS）と呼び，その長さを最長共通部分列長という．ここでは，スクリーン名を X，URL を Y としたときに次式で表される値を利用する．

$$LCSR = \frac{length(LCS(X,Y))}{\max(length(X), length(Y))} \quad (2)$$

(4) **ユーザ名と自己紹介文の共通文字列比（Name-bio-rate）**

POI 公式アカウントではユーザ名に POI 名称を含むことが多く，自己紹介文の中でも POI について説明するため POI 名称が出現しやすい．この傾向を取り入れるため，特徴量としてユーザ名と自己紹介文の LCS を利用する．ただし，ユーザ名は最大 20 文字，自己紹介文は最大 160 文字であり，LCSR を利用すると値が小さくなるため，LCS をユーザ名の長さで正規化した次式を利用する．ここで，X はユーザ名，Y は自己紹介文を表す．

$$\frac{length(LCS(X,Y))}{length(X)} \quad (3)$$

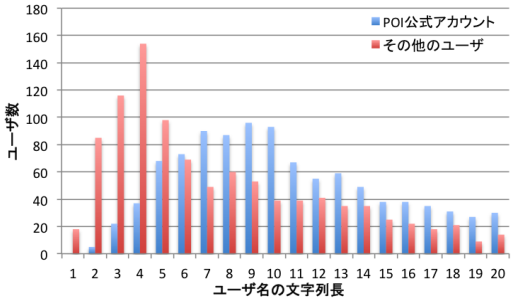



図 2　ユーザ名文字列長の分布
**Fig. 2**　Histogram of user name length.

(5) **ユーザ名の長さ（Name-length）**

POI 名称をユーザ名に含む場合，一般的な人名よりユーザ名が長いと考え利用する．**図 2** に 3 章の評価で用いた POI 公式アカウントとその他のユーザの文字列長の分布を示す．図のように分布が異なることから有効な特徴であると考えられる．

(6) **自己紹介文中での URL 記載有無（URL-bio）**

URL 項目の記載有無と同様の理由であるが，URL 項目ではなく自己紹介文に URL を記載している場合もあるため自己紹介文での記載有無を利用する．特徴量は正規表現により URL 記載有無を判定し 2 値で表す．

(7) **自己紹介文中での電話番号記載有無（Tel-bio）**

一般ユーザはプライバシーの観点から連絡先を記載することは少ないと考えられるが，POI アカウントでは一般ユーザに情報提供することが目的のため連絡先が記載されていることがある．電話番号の記載有無を正規表現により判定し，特徴量は記載有無を 2 値で表す．正規表現の例としては 0¥d{1,4}-¥d{1,4}-¥d{4} のようなパターンを抽出する．

(8) **自己紹介文中でのメールアドレス記載有無（Mail-bio）**

前述の電話番号記載有無と同様の考えで利用する．メールアドレスも電話番号と同様に正規表現により判定し，特徴量としては記載有無を 2 値で表す．

(9) **自己紹介文中での営業時間記載有無（Hours-bio）**

「10:00～12:00」「13:00-20:00」「11 時 30 分から 23 時」などの営業時間の記載は POI 固有の情報と考えられる．そこで，これらを正規表現により抽出し有無を特徴量とする．

(10) **自己紹介文中での郵便番号記載有無（Postal-bio）**

この特徴量も (6)～(9) と同様の考えで利用する．正規表現を使い抽出し，記載有無を 2 値の特徴量とする．

(11) **自己紹介文中での公式記載有無（Official-bio）**

POI に限らず，公式アカウントの中には自己紹介文に公

式であることを明記している場合がある．そのため，自己紹介文に「公式」「オフィシャル」「official」などの記載があるか正規表現により判定し 2 値の特徴量とする．「公式」などを含むかどうかだけを判定すると「非公式」も抽出されるため「非公式」「アンオフィシャル」「unofficial」などを含む場合は記載がないと判定する．

なお，提案手法で用いている POI 固有の特徴量は正規表現で抽出を行える特徴のみを利用しており，国によりパターンの変更がある場合もあるが，同様の表現を抽出できればよいため言語依存性はない特徴量と考えられる．

#### 2.2.2　知名度に関する特徴量

(1) **フォロワ数（Followers）**

Twitter において，企業や有名人の公式アカウントなど情報配信を積極的に行うアカウントではフォロワ数が多い傾向があるためフォロワ数の数値を利用する．

(2) **フレンド数（Friends）**

企業や有名人の公式アカウントではフォロワ数が多く，フレンド数（フォロー数）が少ない傾向があるためこの数値を利用する．

(3) **比率（フォロワ数/フレンド数）（FF-rate）**

全国的に有名な POI であればフォロワ数の絶対値が大きくなるが，ローカルな POI の場合，絶対値は必ずしも大きいとは限らない．しかしながら，フレンド数よりフォロワ数が多いと考えられる．そのため，知名度による影響を勘案しフォロワ数とフレンド数の比率（フォロワ数/フレンド数）を利用する．

(4) **被登録リスト数（Listed-count）**

Twitter にはリストと呼ばれる機能があり，ユーザが特定のユーザをリストに登録することで，登録したユーザの投稿のみをタイムラインで表示できる．そのため，リストに登録されている数（以下，被登録リスト数）はフォロワ数と同様に有名人や情報配信系のアカウント，公式アカウントで数が多いと考えられる．

(5) **認証バッジの有無（Verified）**

認証バッジの有無は，Twitter 社が本人であることを確認しているユーザに付与される[*7]ため採用した．

知名度に関する指標は那須野ら [24] の研究を参考にした．

#### 2.2.3　プロフィール画像特徴量

本研究では画像特徴量として Twitter のプロフィール画像の Bag-of-Visual Words 表現を利用する．Bag-of-Visual Words は画像認識の分野で，特に一般画像認識に利用される画像特徴量 [12] である．本研究では画像上の固定間隔ごとに画像特徴量を計算する dense sampling [13] という手法を利用する．画像特徴量の記述には，SIFT 特徴量 [14] を利用する．SIFT アルゴリズムは特徴点検出と，検出された特徴点周辺の局所的な特徴を記述する特徴量記述の 2 つからなる．Dense sampling では SIFT アルゴリズムの特徴点検出の部分を固定の間隔の画素を対象として特徴量記述を行う．

特徴量記述では，計算対象の画素の周辺との勾配を計算し，勾配方向のヒストグラムを 128 次元のベクトルとして表現する．SIFT 特徴量の詳細は文献 [14] を参照されたい．BoVW では，各画像の特徴ベクトルを k-means クラスタリングなどでベクトル量子化し visual words と呼ばれる特徴ベクトルを生成する．この特徴ベクトルをまとめたものを codebook と呼び，各画像は codebook に含まれる特徴ベクトルの出現頻度のヒストグラムとして表現される．この表現は元々画像認識の分野で，言語処理でよく用いられる文書を単語の集合として扱う Bag-of-words モデルを画像分類に適用し，各画像を visual words の集合として表現したことから Bag-of-Visual Words と呼ばれる．各画像に対して特徴ベクトルを計算する流れを以下に示す．

1. 全画像から SIFT 特徴量を抽出．
2. SIFT 特徴量を k-means クラスタリングでベクトル量子化し codebook を生成．
3. Codebook を元に画像ごとに各特徴ベクトルの出現頻度のヒストグラムを作成し画像の特徴量とする．

本研究では各ユーザのプロフィール画像を BoVW 表現した特徴ベクトルを特徴量として利用する．

### 2.3　機械学習を利用した POI 公式アカウント判定

POI 公式アカウントは 2.1 節，2.2 節で説明した特徴量を利用し，機械学習により判定する．本研究では教師あり機械学習の分類器として Support Vector Machine（以下，SVM）[15] を利用する．分類器の選定基準は，従来研究で用いられていること，一般に広く利用されているライブラリを利用して簡易に実装できるものとした．本研究の主目的は POI 公式アカウント判定を行うために効果のある特徴量を明らかにすることであり，分類器の選択は対象外とした．

## 3.　評価実験

提案手法による POI 公式アカウント分類の性能を評価するため実験を行った．評価は次の 4 つの観点で実施した．1 つめの評価では，1 章で述べた課題 1 の POI 登録数に起因するカバレッジ不足に関する課題が解決されているか確認する（評価 1）．具体的には，POI DB を利用した場合と提案手法を用いた場合で，POI DB に含まれない POI 公式アカウントを含む正解データに対して抽出できる POI 公式アカウントの数を評価した．2 つめの評価では，1 章で述べた課題 2 の登録名称異なりに起因するカバレッジ不足に関する課題が解決されているか確認する（評価 2）．そのため，すべての POI が POI DB に含まれる POI 公式アカウントを対象に，POI DB を利用した場合と提案手法を

[*7] https://support.twitter.com/articles/268350

用いた場合で抽出できるPOI公式アカウントの数を評価した．3つめの評価として，課題3の抽出誤りに関する課題が提案手法では解決されているか確認する（評価3）．具体的には，Twitter API[*8]のユーザ検索機能を利用し，スポット名を検索語として取得したユーザに対してPOI DBを利用した場合と提案手法を用いた場合で，各アカウントに対する分類の正解率を評価した．4つめの評価は，提案手法の総合的な分類性能評価である．従来手法の特徴量および提案する特徴量を含め様々な特徴量の組み合わせのパターンを作り，交差検証によりPOI公式アカウントの分類性能に関する評価を行い，どの組み合わせのパターンが抽出精度が高いか評価を行った（評価4）．

### 3.1　データ

本節では実験に利用するデータについて説明する．実験では2つのデータセットを利用し評価を行った．

#### 3.1.1　データセット1

ここでは，評価1および評価3で利用するデータについて説明する．評価1ではPOI DBに含まれないPOI公式アカウントを取得する必要がある．評価3ではスポット名でアカウントを検索した結果が必要となる．そのため，Twitter APIでアカウント検索を行いデータを取得した．本研究では外出時にローカル検索を行った場合に公式アカウントが情報源として有用であると想定しているため，Foursquare（Swarm）[*9]のチェックイン数が多いスポット名をTwitter APIの検索語としてアカウントを検索した．検索結果に対して人手でPOI公式アカウントかどうか，およびFoursquareへのPOI登録有無を確認したものを正解データとした．たとえば，「名古屋駅」で検索した場合，検索結果として「LEC名古屋駅前本校」のTwitterアカウントが取得でき，これはFoursquareにはPOIとして登録されていないがPOI公式アカウントとなる．検索語に利用したスポット名は，Foursquareから日本全国のPOI 966,614件を取得しチェックインユーザ数の上位に含まれるスポットのうち駅や空港など交通機関のスポット上位15個と，それ以外のスポット上位19個の合計34個を利用した．交通機関とそれ以外で分けているのは，Foursquareでは交通機関へのチェックインが多く，上位100位までのうち74件が駅または空港であったため，スポットのカテゴリのバランスを取るためである．上記手順で作成した正解データ数を**表2**に示す．また，Twitter APIで検索語に利用したスポット名と検索結果のアカウント数を付録表A·1に示す．検索時は各検索語での最大取得数を20アカウントとした．

#### 3.1.2　データセット2

2つめのデータセットは評価2および評価4で利用するデータセットである．評価2ではPOI DBに含まれるPOI公式アカウントのデータが必要となる．また評価4では分類器の学習データが必要になる．Swarmから投稿されたツイートでは，一部のツイートでPOIと公式アカウントが関連付けられているため，それを正例の元データとして利用する．負例はランダムサンプリングしたユーザを利用する．正例，負例それぞれのデータ作成手順の詳細を以下に示す．

**表2**　データセット1の各事例数
**Table 2**　Number of positive and negative samples.

| 種別 | | データ数 |
|---|---|---|
| 正例 | POI DBに登録あり | 139 |
| | POI DBに登録なし | 28 |
| 負例 | | 203 |

**正例データ作成手順**

(1) **POIに関連付けられたスクリーン名の一覧を取得**

Swarmから投稿されたツイートではFoursquare社がPOIとTwitterのアカウントを対応付けている場合，以下のようなツイートになる．

- I'm at 横浜赤レンガ倉庫 イベント広場-@yokohama-redbric in 横浜市，神奈川県 https://www.swarmapp.com/xxxx
- 雨やし地下から到着（ノ・∀・）ノ (@あべのハルカス (ABENO HARUKAS) - @abenoharukas in 大阪市，大阪府) https://www.swarmapp.com/xxxx

そこで，投稿元がFoursquareのツイートから正規表現でスクリーン名（@xxxxの名称）を抽出した．データの取得元の期間は2015年7月である．なお，Foursquareに存在するPOIのすべてに各POIの公式Twitterアカウントが関連付けられている場合，本研究の手法が不要になるが，筆者らの調査（付録A.2参照）ではPOIの公式Twitterアカウントが関連付けられているのは約16.3%のPOIであった．そのためFoursquareのすべてのPOIが公式アカウントと関連付けられているわけではない．

(2) **Twitter APIでユーザのプロフィール情報を取得**

上記(1)で取得できた4,251個のアカウントに対してスクリーン名をクエリとしてTwitter APIを利用しユーザ名，自己紹介文，フォロワ数，フレンド数などのプロフィール情報を取得した．取得したアカウントに対して自己紹介文や投稿内容から非公式アカウントを人手で削除した．(1)で取得したSwarm経由のツイートを機械的に正例とせず，目視確認を行ったのは，Swarm上で関連付けられているTwitterアカウントの中には非公式アカウントも含まれており，非公式アカウントは正例から削除するためである．

(3) **条件によるフィルタリング**

(2)までに得られたユーザに対して，次の4つの条件でフィルタリングを行った．

条件1．言語設定が日本語

[*8] https://dev.twitter.com/rest/public
[*9] https://www.swarmapp.com/

条件 2．海外やオンラインショップのアカウントを除外
条件 3．プロフィール画像が取得可能
条件 4．2015 年 1 月〜7 月の投稿数が 20 以上

条件 1，2 は本研究では日本における訪問可能な POI を対象とするための条件である．条件 3 はデータ作成中にユーザによりプロフィール画像が変更される場合があったため含めた条件である．条件 4 は本研究では機械学習の特徴量として 2.1.1 節で説明した投稿文章を利用するため一定量のツイートが必要になるための条件である．

以上の手順で作成した 1,000 アカウントを正例のデータとした．

**負例データ作成手順**

(1) **非 POI 公式アカウント候補を取得**

正例と同じ時期の非 POI 公式アカウントを抽出するため，2015 年 7 月 1 日〜10 日に投稿された日本語ツイートから各日 200 ツイートずつランダムサンプリングし 2,000 アカウント抽出した．

(2) **条件によるフィルタリング**

上記で得られたユーザに対して，次の 4 つの条件でフィルタリングを行った．

条件 1．スパムユーザ，bot を除外
条件 2．POI 公式アカウントを除外
条件 3．プロフィール画像が取得可能
条件 4．2015 年 1 月〜7 月の投稿数が 20 以上

スパム [27] や bot [26] などの自動投稿のアカウントは既存研究で判定できるため除外した．条件 2 では (1) でランダムサンプリングしたアカウントには POI 公式アカウントが含まれてしまうため手動で除外した．条件 3，4 については正例と同様である．

課題 2（POI DB との登録名称の異なり）に該当するデータは，正例のデータ 1,000 件中 567 件である．

本研究では，先行研究 [1], [2], [4] と同様に特徴量の効果を見るため正例と負例のデータ数を同数とした．プロフィール画像はユーザがアップロードした画像を Twitter 社が 48 × 48 に加工した normal の画像[*10]を利用し dense sampling のため 2 ピクセルごとに SIFT 特徴量を算出した．評価 4 では前述のデータを用いて 10 分割交差検定を行った．

### 3.2　実験環境

SVM の実装には Python 2.7 および scikit-learn 0.16 を利用し，カーネルは線形カーネルを利用した．SIFT 特徴量の計算には画像処理ライブラリ OpenCV 2.4.11 を用いた．実験には Intel(R) Xeon(R) CPU E5-2130 v2@2.6 GHz，メモリ 64 GB のマシンを利用した．SVM のパラメータは，グリッドサーチを行い最も性能がよいパラメータ（C = 1）を利用した．BoVW の特徴ベクトルの次元（codebook サイズ）は k-means クラスタリングのクラスタ数によって決まるため，クラスタ数を変えて実験を行い k = 100 とした．形態素解析器には MeCab[*11] を利用した．

### 3.3　POI DB との比較評価結果と考察（評価 1，2）

本節では，POI DB を利用した場合と提案手法を用いた場合での，抽出できる POI 公式アカウント数の評価について説明する．評価 1 では POI DB に含まれない POI も対象とし，評価 2 では POI DB に含まれる POI のみを対象とする．評価指標には確信度 [25] と抽出数の比を利用する．

$$\text{確信度} = \frac{|X \cap Y|}{|X|} \quad \text{抽出数比} = \frac{|Y|}{|X|}$$

ここで，$X$ は POI DB を利用した場合に抽出された POI 公式アカウントの集合，$Y$ は提案手法で抽出された POI 公式アカウントの集合を示す．確信度を評価指標とする理由は，提案手法で判定できる POI 公式アカウントが，ナイーブな手法である POI DB を用いて名称の一致により抽出した結果をどれだけ含んでいるか評価するためである．抽出数の比を評価指標として用いる理由は，確信度の値だけでは各手法で抽出できるアカウントが同じ場合でも値が大きくなり，手法の優劣が判断できないため抽出できる総数で比較を行うためである．確信度が高いと POI DB を用いた場合に抽出できる POI 公式アカウントを提案手法でも抽出できていることを示す．一方，抽出数の比が 1 より大きいほど提案手法がより多く POI 公式アカウントを抽出できたといえる．そのため，確信度が高く，抽出数の比が大きいほど提案手法がよいといえる．

**ベースライン**　ユーザ名に POI 名称を含み，かつ自己紹介文に公式であることが記載されているユーザを抽出したものをベースラインとする．POI DB として，評価 1 では 3.1.1 項で説明した日本全国の POI 966,614 件を利用し，評価 2 では 3.1.2 項で説明した正例の学習データを作成した期間と同じ 2015 年 7 月に Foursquare を経由した投稿に含まれる 18,339 件の POI を利用した．Foursquare 経由の投稿では 3.1.2 項で示したようなツイートになるため，POI 名称をルールにより抽出でき，これを POI DB とした．そのため，評価 2 では POI DB には抽出対象のすべての POI が登録されている状態となる．

ベースライン手法での適合率と再現率は，評価 1 はそれぞれ 0.787，0.222，評価 2 ではそれぞれ 0.987，0.304 であった．なお，適合率は POI 公式アカウントと判定したユーザのうち人手で正解ラベルを付与したユーザの割合，再現率は人手で POI 公式アカウントとラベル付けしたユーザのうち，いくつのユーザを抽出できたかという割合である．

提案手法では投稿文章の BoW，自己紹介文の BoW，POI

[*10] https://dev.twitter.com/overview/general/user-profile-images-and-banners

[*11] http://taku910.github.io/mecab/

表 3　抽出数比較評価結果
**Table 3**　Comparison of the number of extracted accounts.

| | ベースライン (X) | 提案手法 (Y) | 両方で抽出された数(X∩Y) | 確信度 | 抽出数比 |
|---|---|---|---|---|---|
| 評価 1 | 37 | 152 | 37 | 1 | 4.1 |
| 評価 2 | 301 | 933 | 286 | 0.95 | 3.1 |

表 4　評価 1 の抽出結果
**Table 4**　Result of Evaluation 1.

| 種別 | ベースライン | 提案手法 |
|---|---|---|
| POI DB 登録あり | 35（25.2%） | 133（95.7%） |
| POI DB 登録なし | 2（7.14%） | 19（67.9%） |
| 合計 | 37（22.2%） | 152（91.0%） |

固有特徴量，知名度に関する特徴量，画像特徴量を利用してアカウントを分類した．**表 3** に評価結果を示す．表 3 の評価 1 の結果から提案手法では，POI DB を用いた場合に抽出される POI 公式アカウントを 100%カバーし，抽出される POI 公式アカウントの総数は 4.1 倍であった．**表 4** に POI DB に含まれる公式アカウントと含まれない公式アカウントの個別の抽出結果を示す．表中の値は各種別での抽出結果数を示し，括弧内はカバー率を示す．POI DB に登録があるものに対してはベースラインより 3.8 倍，POI DB に登録がない POI に関しては 9.5 倍のカバー率となった．この結果より，POI DB に含まれない POI 公式アカウントに対しても提案手法が有効であることを確認した．次に，表 3 の評価 2 の結果について考察する．表 3 から提案手法では，POI DB を用いた場合に抽出される POI 公式アカウントのうち 95%をカバーし，抽出される POI 公式アカウントの総数は 3.1 倍であった．そのため，POI DB を利用した名称の一致による手法と比較して機械学習を用いる提案手法のカバー率が高いことが示された．また，3.1.2 項で述べたとおり，課題 2（POI DB との登録名称の異なり）に該当するデータは，データセット 2 の正例データ 1,000 件中 567 件であり，このうち 528 件（93.1%）を提案手法により抽出できており，課題 2 を解決できたと考えられる．

### 3.4　抽出誤りに関する評価結果と考察（評価 3）

データセット 1 を対象に，データセット 2 の 2,000 件のデータから分類器を学習したモデルを利用し評価を行った．抽出誤りに対しては正例負例どちらの判定も正確に行えることが望ましいため評価指標には下記の正解率を検索語にわたって平均した平均正解率を用いる．

$$\text{正解率} = \frac{\text{予測ラベル正解率}}{\text{サンプル数}}$$

ベースラインには，評価 1，2 と同様に POI DB を利用しユーザ名に POI 名称を含むユーザを公式アカウントと判定する方法を用いた．付録 A.1 に示した検索語ごとに正解率を算出し全体で平均をとった平均正解率の結果は，ベースライン 0.659，提案手法 0.823 となった．この結果に対して t 検定を実施したところ有意水準 5%で有意な差があった．

表 5　特徴量の次元数
**Table 5**　Feature dimension.

| 特徴量 | 次元数 |
|---|---|
| 投稿文章 BoW | 381507 |
| 自己紹介文 BoW | 14098 |
| POI 固有特徴量 | 11 |
| 知名度に関する特徴量 | 5 |
| 画像特徴量 | 100 |

### 3.5　分類性能に関する評価結果と考察（評価 4）

分類性能の定量的な評価指標には適合率，再現率，F 値を用いる．適合率および再現率は 3.3 節で説明したとおりである．F 値は適合率と再現率の調和平均であり次式で計算する．

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

次に，分類性能評価で比較対象とする 3 つのベースライン手法について説明する．

**ベースライン 1（B1）**　ユーザの投稿文章の BoW を特徴量に利用し，SVM で分類したものをベースラインとする．これは先行研究 [4], [7] と同様の手法である．

**ベースライン 2（B2）**　ユーザの自己紹介文の BoW を特徴量に利用し，SVM で分類したものをベースラインとする．先行研究 [7] と同様の手法である．

**ベースライン 3（B3）**　ユーザの投稿文章および自己紹介文の BoW を特徴量に利用し，SVM で分類したものをベースラインとする．先行研究 [7] と同様の手法である．

SVM の入力となる特徴量の次元数を**表 5** に示す．評価では特徴量の組み合わせを変えて実験を行った．評価結果を**表 6** に示す．太字は最も性能が良かったことを示す．表中の B1 から B3 は前述のベースラインを示す．POI 固有+知名度は POI 固有特徴量と知名度に関する特徴量を並列に並べたものを特徴量として判定した場合を示し，同様に，POI 固有+知名度+画像は提案手法する特徴量すべてを用いた場合を示す．また，B3+POI 固有は，ユーザの投稿文章および自己紹介文の BoW と POI 固有特徴量を並列に並べたものを特徴量とした場合を示す．その他の結果についても同様である．知名度に関する特徴量だけを利用する場合，POI に限らず公式アカウントを判定することになるため，必ず POI 固有特徴量との組み合わせで利用した．表中の † はベースライン 3 と比べて有意水準 5%で差があることを，‡ は有意水準 1%で差があることを示す．B3+POI 固有を用いた場合，ベースライン 3 と比べ，再現率は有意

表 6　性能評価結果
**Table 6**　Result of performance evaluation.

| 手法 | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| B1 | 0.919 | 0.916 | 0.917 |
| B2 | 0.962 | 0.724 | 0.826 |
| B3 | 0.938 | 0.918 | 0.928 |
| POI 固有 | 0.862 | 0.879 | 0.87 |
| POI 固有+知名度 | 0.878 | 0.885 | 0.881 |
| 画像 | 0.728 | 0.706 | 0.716 |
| POI 固有+知名度+画像 | 0.883 | 0.886 | 0.884 |
| B3+POI 固有 | 0.941 | 0.924‡ | 0.932† |
| B3+POI 固有+知名度 | 0.941 | 0.922 | 0.931 |
| B3+画像 | 0.939 | 0.924 | 0.931 |
| B3+POI 固有+知名度+画像 | 0.944 | **0.933**† | **0.938**† |

水準 1%で有意な差があり，F 値は有意水準 5%で有意な差があり性能を改善した．また，B3+POI 固有+知名度+画像を用いた場合，ベースライン 3 と比べ，再現率と F 値について有意水準 5%で有意な差があり性能を改善した．なお，有意差検定には交差検定の各試行の結果を利用し，対応のある t 検定を実施した．性能改善の幅に関しては，提案する特徴量をすべて加えた場合が最も改善するという結果となった．適合率についても，提案手法が最も高いという結果であったが有意差はなかった．提案する特徴量のみを用いた場合は，POI 固有のみを用いた場合に比べ，知名度，画像特徴量を加えるに従い性能が改善している．このことから各特徴はアカウントの判定に寄与していると考えられる．

提案する特徴量の各要素の寄与を確認するため，提案する特徴量のみを用いた場合の各特徴の重みを**表 7** に示す．表 7 (a) は POI 固有特徴量と知名度に関する特徴量のみを用いた場合の重みを降順に示し，表 7 (b) はさらに画像特徴量も加えた場合の重みの上位 15 件を降順に示している．どちらの場合も POI 固有特徴量と知名度に関する特徴量で効果がある特徴量は同様の傾向があり Location および URL 項目の記載有無，自己紹介文での電話番号，営業時間の重みが大きくなっている．これらの特徴は POI であることを表すため効果があったと考えられる．自己紹介文での公式の記載有無は公式判定に効果があると考えられる．また，スクリーン名と URL，およびユーザ名と自己紹介文の共通文字列に関しても重みがあり効果があることが分かる．一方，知名度に関する特徴量に関して，フォロワ数とフレンド数の比率が特徴量全体の中で最も重みが大きくなっている．これは，有名人や全国区の企業の公式アカウントではフォロワ数の絶対値が大きくなるが，POI 公式アカウントではローカルな POI も含まれるため，絶対値より相対値である比率が効いていると考えられる．また，被登録リスト数は重みが大きく，公式アカウントを一般ユーザ

表 7　提案する特徴量の重み
**Table 7**　Weight of feature vectors.

(a) POI 固有+知名度　(b)POI 固有+知名度+画像
(a) POI info + popularity　(b) POI info + popularity + image

| 特徴量 | 重み |
|---|---|
| FF-rate | 2.193 |
| Listed-count | 0.506 |
| Hours-bio | 0.387 |
| Official-bio | 0.353 |
| Tel-bio | 0.313 |
| Name-bio-rate | 0.274 |
| URL-field | 0.263 |
| Loc-field | 0.242 |
| Screen-URL-rate | 0.186 |
| Name-length | 0.152 |
| Mail-bio | 0.009 |
| Postal-bio | -0.0001 |
| Verified | -0.008 |
| URL-bio | -0.029 |
| Followers | -0.089 |
| Friends | -0.158 |

| 特徴量 | 重み |
|---|---|
| FF-rate | 2.465 |
| Listed-count | 0.582 |
| Hours-bio | 0.455 |
| Tel-bio | 0.421 |
| Official-bio | 0.415 |
| Loc-field | 0.316 |
| Name-bio-rate | 0.307 |
| URL-field | 0.281 |
| Screen-URL-rate | 0.212 |
| 画像特徴量 No.46 | 0.17 |
| 画像特徴量 No.97 | 0.131 |
| Name-length | 0.129 |
| 画像特徴量 No.14 | 0.11 |
| 画像特徴量 No.86 | 0.106 |
| 画像特徴量 No.41 | 0.103 |

のタイムラインとは区別して閲覧しているユーザがいることが分かる．認証バッジの有無は重みが非常に小さくなっている．これは認証バッジがついている場合は公式であるが，必ずしも POI に関係するわけではないため判定には寄与しにくいものと考えられる．

## 4. 関連研究

関連する研究として，Twitter ユーザの属性推定の研究と Web やソーシャルメディアからの POI 抽出の研究がある．

### 4.1 Twitter ユーザ属性推定の研究

Twitter ユーザの属性推定の研究が数多く取り組まれている．属性推定の研究は（1）投稿内容に基づく手法，(2) ソーシャルグラフに基づく手法に分けられる．

#### 4.1.1 投稿内容に基づくユーザ属性推定

Rao ら [1] は顔文字や略語などの社会言語学に基づく特徴と N-gram を特徴量として，SVM により年齢，性別，地域，政治的志向を推定する手法を提案した．池田ら [4] は属性ごとの特徴語を赤池情報量基準（AIC）により選択し，特徴語を特徴量として SVM で性別，年代，居住地域の推定を行った．平野ら [6] は属性の相互依存関係を考慮するため Markov Logic を用いて複数の属性を同時に考慮しながら集合的に推定する手法を提案した．たとえば，ユーザが高校生と推定されたときは，そのユーザの年代が 10 代である可能性が高いなどの依存関係を考慮している．伊藤ら [5] は Twitter と Blog の共通ユーザを用いて Blog に記載され

ているプロフィールを学習データとしてTwitterユーザのプロフィールを推定した．Blogではプロフィール欄が自由記述ではないことを利用し教師ラベルを自動的に獲得した．榊ら [7] は投稿内容，自己紹介文，ユーザが含まれるリスト名からSVMを利用して対象ユーザの職業を推定した．その他，居住地 [3], [16] や発言位置の推定 [17], [18] を行う研究がある．

#### 4.1.2　ソーシャルグラフに基づくユーザ属性推定

ソーシャルグラフに基づくユーザ属性推定の研究では，つながりのあるユーザは互いに似た属性を持つと仮定してユーザ属性を推定する．Pennacchiotti ら [2] はTwitterのプロフィール文書，返信や投稿方法などのツイートの仕方，ツイートによく現れる単語，友人の数などソーシャルネットワークの特徴を元に政治的所属，民族，特定のビジネスへの興味を推定した．またソーシャルグラフをもとにラベル更新を行う手法を提案した．上里ら [8] は属性推定対象ユーザとフォローやメンションの関係がある周辺ユーザに対しても属性推定をあらかじめ適用することで推定対象ユーザの推定精度を向上させる手法を提案した．

投稿内容に基づく手法もソーシャルグラフに基づく手法も，従来は一般ユーザを対象とした属性推定の研究が行われており，従来研究をPOI公式アカウント判定に適用した場合の定量的評価は行われていない．また，POI固有の特徴量を利用して判定を行った研究はない．

### 4.2　WebやソーシャルメディアからのPOI抽出の研究

倉島ら [19] はFlikcrにおけるジオタグ付き写真からMean-Shiftクラスタリングにより人気の観光スポットを抽出した．スポット名称にはFlickrで投稿されたタグを利用した．荒川ら [20] は倉島らと同様にMean-Shiftクラスタリングにより人気の観光スポットを抽出し，Foursquareなどのチェックインサービスと統合することにより正確なPOI名称を得られることを明らかにした．Rae ら [21] はWeb検索のスニペットからPOIを抽出する研究を行った．この研究ではWeb検索のスニペットからCRFを使ってPOI名称を抽出し，場所の特定はFlickrのジオタグ付き写真のタグを使い，1 km四方でタグの出現確率を計算した．本研究で提案するPOI公式アカウント判定を行い，その後場所の特定まで行った場合，TwitterからPOIを抽出していると見なすことができる．従来研究では，WebやFlickrデータからPOIを抽出しており，Twitterなどの SNSのプロフィール情報からPOIを抽出する研究はない．

## 5.　おわりに

本稿ではTwitterの投稿文章とプロフィール情報（自己紹介文，プロフィール画像）を用いてPOI公式アカウントを分類する手法を提案した．従来手法で用いられていた投稿文章と自己紹介文のBoWに加え，POI固有特徴量，知名度に関する特徴量，プロフィール画像から作成した画像特徴量を提案した．POI DBを利用しユーザ名とPOI名称の一致によりPOI公式アカウントを判定する手法と提案手法のカバー率を比較し，提案手法の方が約3倍程度抽出数が多いことを確認した．投稿文章と自己紹介文のBoWに提案する特徴量を追加することで，従来手法より再現率，F値を向上できることを定量評価により示した．提案手法では再現率0.933，F値0.938の性能で分類できることを確認した．

今後の課題は，Rao ら [1] のように特徴量ごとに分類器を作成し，その結果に対してさらに分類を行うStacked Modelの検討や，計算時間短縮のため特徴量の数が多いBoWを利用せずPOI固有特徴量を拡充したり画像特徴量の色情報を利用したりするなどが考えられる．

## 参考文献


[1] Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M.: Classifying Latent User Attributes in Twitter, *Proc. 2nd International Workshop on Search and Mining User-generated Contents*, *SMUC '10*, pp.37–44 (2010).

[2] Pennacchiotti, M. and Popescu, A.M.: Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter, *KDD '11*, pp.430–438 (2011).

[3] Hecht, B., Hong, L., Suh, B. and Chi, E.H.: Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles, *Proc. SIGCHI Conference on Human Factors in Computing Systems*, *CHI '11*, pp.237–246 (2011).

[4] 池田和史，服部　元，松本一則，小野智弘，東野輝夫：マーケット分析のためのTwitter投稿者プロフィール推定手法，情報処理学会論文誌コンシューマ・デバイス&システム（CDS），Vol.2, No.1, pp.82–93 (2012).

[5] 伊藤　淳，西田京介，星出高秀，戸田浩之，内山匡：TwitterとBlogの共通ユーザおよび会話ユーザの同類性に着目したTwitterユーザ属性推定，日本データベース学会論文誌，Vol.12, No.1, pp.31–36 (2013).

[6] 平野　徹，牧野俊朗，松尾義博：Markov Logicを用いたテキストからのユーザ属性推定，Vol.27，人工知能学会，pp.1–4 (2013).

[7] 榊　剛史，松尾　豊：ソーシャルメディアユーザの職業推定手法の提案，知能と情報，Vol.26, No.4, pp.773–780 (2014).

[8] 上里和也，浅井洋樹，奥野峻弥，山名早人：Twitterユーザを対象とした属性推定の精度向上—周辺ユーザの属性補完を利用して，第7回データ工学と情報マネジメントに関するフォーラム（DEIM 2015），pp.D8–5 (2015).

[9] 高村大也：言語処理のための機械学習入門（自然言語処理シリーズ），コロナ社 (2010).

[10] Manning, C.D., Raghavan, P. and Schütze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008).

[11] Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *Workshop on statistical learning in computer vision, ECCV*, Vol.1, No.1-22, pp.1–2 (2004).

[12] 八木康史，斎藤英雄（編）：CVIMチュートリアルシリーズ コンピュータビジョン最先端ガイド3，アドコム・メディア (2010).

[13] Jurie, F. and Triggs, B.: Creating efficient codebooks for

visual recognition, *10th IEEE International Conference on Computer Vision, 2005, ICCV 2005*, Vol.1, IEEE, pp.604–610 (2005).
[14] Lowe, D.G.: Distinctive image features from scale-invariant keypoints, *International journal of computer vision*, Vol.60, No.2, pp.91–110 (2004).
[15] Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, Vol.20, No.3, pp.273–297 (1995).
[16] Cheng, Z., Caverlee, J. and Lee, K.: You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users, *ACM CIKM '10*, pp.759–768 (2010).
[17] 山口祐人，伊川洋平，天笠俊之，博之北川：ソーシャルメディアにおけるローカルイベントを用いたユーザ位置推定手法，情報処理学会論文誌データベース（TOD），Vol.6, No.5, pp.23–37 (2013).
[18] 伊川洋平，榎　美紀，立堀道昭：マイクロブログのメッセージを用いた発信場所推定，第 4 回データ工学と情報マネジメントに関するフォーラム（DEIM 2012），pp.F7–2 (2012).
[19] Kurashima, T., Iwata, T., Irie, G. and Fujimura, K.: Travel Route Recommendation Using Geotags in Photo Sharing Sites, *Proc. 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, pp.579–588 (2010).
[20] 荒川　豊，タチアーナシェフラー，ステファンバウマン，アンドレアスデンゲル：ソーシャル観光マップ—ソーシャルデータからの観光スポット抽出，情報処理学会論文誌コンシューマ・デバイス&システム（CDS），Vol.4, No.1, pp.1–11 (2014).
[21] Rae, A., Murdock, V., Popescu, A. and Bouchard, H.: Mining the Web for Points of Interest, *Proc. 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12*, pp.711–720 (2012).
[22] Melamed, I.D.: Bitext Maps and Alignment via Pattern Recognition, *Computational Linguistics*, Vol.25, pp.107–130 (1999).
[23] Tominaga, T. and Hijikata, Y.: Study on the Relationship between Profile Images and User Behaviors on Twitter, *WWW '15 Companion Proceedings of the 24th International Conference on World Wide Web*, pp.825–828 (2015).
[24] 那須野薫，奥山晶二郎，中西鏡子，松尾　豊：Twitter における候補者の選挙地盤に着目した国政選挙の当選者予測，情報処理学会論文誌，Vol.56, No.10, pp.2044–2053 (2015).
[25] 金明　哲：R によるデータサイエンス—データ解析の基礎から最新手法まで，森北出版 (2007).
[26] 蔵内雄貴，西田京介，川中　翔，星出高秀，内山　匡：ベンフォードの法則を応用した bot アカウント検出，日本データベース学会論文誌，Vol.12, No.1, pp.19–24 (2013).
[27] McCord, M. and Chuah, M.: Spam detection on twitter using traditional classifiers, *Proc. 8th international conference on Autonomic and trusted computing, ATC '11*, pp.175–186 (2011).

# 付　　録

## A.1　検索語に利用したスポット名

**表 A·1**　検索語に利用したスポット名と Twitter でのアカウント検索結果数
**Table A·1**　Number of Twitter search result and query for Twitter search.

(a) 交通機関

| スポット名 | 検索結果数 |
|---|---|
| 成田国際空港 | 2 |
| 東京駅 | 9 |
| 新宿駅 | 18 |
| 渋谷駅 | 15 |
| 東京国際空港 | 2 |
| 秋葉原駅 | 13 |
| 京都駅 | 17 |
| 横浜駅 | 15 |
| 新大阪駅 | 3 |
| 池袋駅 | 16 |
| 名古屋駅 | 13 |
| 品川駅 | 11 |
| 上野駅 | 9 |
| 大阪駅 | 11 |
| 関西国際空港 | 2 |

(b) 交通機関以外のスポット

| スポット名 | 検索結果数 |
|---|---|
| 東京国際展示場 | 5 |
| 東京スカイツリー | 15 |
| 幕張メッセ | 18 |
| 東京ディズニーシー | 11 |
| 東京ディズニーランド | 6 |
| 六本木ヒルズ | 11 |
| 東京ドーム | 13 |
| 東京タワー | 16 |
| 日本武道館 | 20 |
| Apple Store 銀座 | 1 |
| ユニバーサル・スタジオ・ジャパン | 6 |
| さいたまスーパーアリーナ | 16 |
| 東京ミッドタウン | 8 |
| パシフィコ横浜 | 16 |
| 清水寺 | 13 |
| ダイバーシティ東京プラザ | 10 |
| 浅草寺 | 9 |
| 東京国際フォーラム | 16 |
| 伏見稲荷大社 | 4 |

## A.2　Foursquare から取得できる POI 公式アカウント

3 章の評価実験において正例の学習データを，Swarm を経由して投稿されたツイートから作成している．Swarm に存在する POI にすべて公式アカウントが関連付けられている場合，本研究の手法が不要になるため関連付けられているツイート数を調査した．**図 A·1** に 2015 年 7 月に Swarm 経由で投稿されたツイート数と，公式アカウントが関連付けられているツイート数を示す．1 カ月分の平均で約 16.3%のツイートが公式アカウントに関連付けられていた．残りの POI に関しては公式アカウントと関連付けが行われていないため，POI と公式アカウントを関連付ける必要がある．

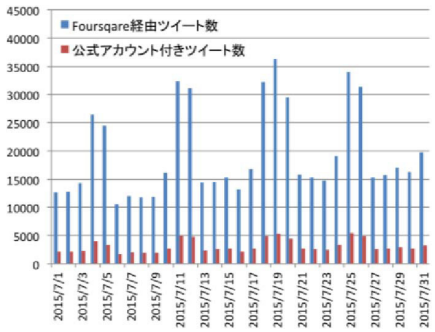
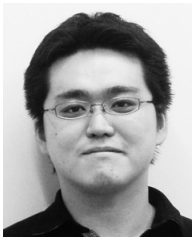



図 A·1　Foursquare 経由の総ツイート数と公式アカウント関連ツイート数

**Fig. A·1**　Total number of tweets from foursquare and POI-associated ones.

**落合 桂一**　（正会員）

2006 年千葉大学工学部情報画像工学科卒業．2008 年同大学院博士前期課程修了．同年株式会社 NTT ドコモ入社．SNS および位置情報解析の研究開発に従事．日本データベース学会会員．

**山田 渉**　（正会員）

2010 年東京理科大学理工学部経営工学科卒業．2012 年東京大学学際情報学府総合分析情報学コース修了．同年株式会社 NTT ドコモ入社．SNS および位置情報解析，ユーザインタフェースに関する研究開発に従事．ACM 会員．

**深澤 佑介**　（正会員）

2002 年東京大学工学部卒業．2004 年東京大学大学院工学系研究科修士課程修了．同年株式会社 NTT ドコモ入社．2011 年東京大学大学院工学系研究科博士後期課程修了．同年 10 月より東京大学人工物工学研究センターにて協力研究員兼任．IEEE，人工知能学会各会員．博士（工学）．

**菊地 悠**　（正会員）

2000 年東京大学精密機械工学科卒業．2002 年同大学院博士前期課程修了．同年株式会社 NTT ドコモ入社．SNS および位置情報解析の研究開発に従事．

**松尾 豊**　（正会員）

1997 年東京大学工学部電子情報工学科卒業．2002 年同大学院博士課程修了．博士（工学）．同年より，産業技術総合研究所研究員．2005 年 10 月よりスタンフォード大学客員研究員．2007 年 10 月より，東京大学大学院工学系研究科総合研究機構/知の構造化センター/技術経営戦略学専攻准教授．人工知能学会編集委員長．専門は，Web マイニング，人工知能，ビッグデータ分析．


（担当編集委員　岡本 昌之）