

研究論文

# 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

池田 和史<sup>1,a)</sup> 服部 元<sup>1</sup> 松本 一則<sup>1</sup> 小野 智弘<sup>1</sup> 東野 輝夫<sup>2</sup>

受付日 2011年5月20日, 採録日 2011年12月2日

**概要:** 近年, Twitter のようなブログや Web 掲示板などに投稿された商品やテレビ番組などに対する口コミ情報を分析してマーケティングなどに応用する評判解析技術に注目が集まっている. これらは手軽に情報発信することが可能なため, 新鮮かつ多数の意見を即座に収集するツールとして, その活用は大きな可能性を持っている. 一方で, 評判は投稿者の年齢や性別, 趣味などのプロフィールに応じて異なることが多いが, ブログや掲示板には投稿者の年齢や性別が記載されていない場合が多く, 投稿数や平均的な意見などの表面的な情報しか抽出できず, プロフィールごとの意見を抽出できないことが課題であった. この問題を解決するため, 著者らは Twitter 上の口コミ投稿者の日常的な投稿内容を解析することで, 年代, 性別, 居住地域などのプロフィールを推定する技術を開発した. 本技術を利用することで, ネット上の口コミ情報をプロフィールごとに分類, 集約することが可能となり, 商品の改善やテレビ番組の企画などに生かすことが可能となる. 性能評価実験の結果, 提案手法の汎用的な推定精度は性別で 88.0%, 年代で 68.0%, 居住地域で 70.8%であり, 視聴率測定などへの応用を想定したプロフィール分布誤差の評価では, 分布に偏りがある場合でも性別で 8.8%, 年代で 12.4%, 居住地で 14.0%と実利用に十分な精度であることが示された.

**キーワード:** Web マイニング, 評判解析, プロフィール, Twitter, アプリケーション

## Demographic Estimation of Twitter Users for Marketing Analysis

KAZUSHI IKEDA<sup>1,a)</sup> GEN HATTORI<sup>1</sup> KAZUNORI MATSUMOTO<sup>1</sup> CHIHIRO ONO<sup>1</sup> TERUO HIGASHINO<sup>2</sup>

Received: May 20, 2011, Accepted: December 2, 2011

**Abstract:** This paper proposes a real-time analysis technology of the online opinions of commercial products and broadcast TV programs. As many people submit their opinions via social media services, such as Twitter, utilizing these real-time and huge amounts of opinions is strongly desired as a novel marketing tool. However, it is impossible in many cases to understand the overall trend of such enormous user opinions by browsing the information stream on the screen. In addition, though presuming the ratio of positive and negative opinions is useful, that discrimination is not much enough because the ratio of opinions differs depending on user demographics (age, sex, area, etc.) The proposed technology makes it possible to analyze the contents of Twitter streams related to commercial products or broadcast TV programs, and estimate the demographics of the users by tracking and analyzing their past tweets. This analysis attracts peoples such as, product planners, broadcast TV directors, and advertisement agencies that produce and promote products/TV programs for target segments. Our experimental results show that the estimation accuracy of the proposed algorithms is, 88.0% in sex, 68.0% in age, 70.8% in areas, respectively. The error ratio in the distribution of estimated demographics was 8.8% in sex, 12.4% in age, 14.0% in area, respectively, which is high enough for practical use.

**Keywords:** Web mining, sentiment analysis, profiling, Twitter, application

<sup>1</sup> KDDI 研究所  
KDDI R&D Laboratories Inc., Fujimino, Saitama 356-8502, Japan

<sup>2</sup> 大阪大学大学院情報科学研究科

Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan

a) kz-ikeda@kddilabs.jp

### 1. 研究の背景

近年、インターネットの普及により、ブログや Web 掲示板などを通して一般ユーザが情報を発信、閲覧することでコミュニケーションを行う、ソーシャルネットワーキングサービス (SNS) が普及している。特に最近では、個々のユーザが「つぶやき」と呼ばれる短文をネット上に投稿し、閲覧・共有することが可能な Twitter [1] が急速に普及し、2010 年 9 月には国内大手 SNS である mixi [2] を上回り、1200 万ユーザを超えたといわれる [3]。

これらの SNS 上では、商品やコンテンツに対する感想や意見などの口コミ情報も投稿されており (図 1)、これらを分析してマーケティングに応用する評判解析技術 [4] に注目が集まっている。従来のマーケティングでは、アンケートによるモニタ調査が主流であったのに対し、ネット上の口コミ情報を利用することで、大量の意見をリアルタイムに、低コストで調査することが可能となる。一方、商品やコンテンツに対する口コミは年齢や性別、居住地域などのプロフィールに応じて異なる。たとえば、若者にとって「面白い」と好評のテレビ番組が年配者にとっては「騒々しい」などの悪評を得る場合がある。このため、マーケティ

ングにおいては、(1) 商品やコンテンツに対して言及している投稿者のプロフィールの分布傾向を知りたい、(2) 特定のプロフィールを持つ投稿者の意見を収集して分析したい、といった需要が大きい。ところが、ブログや掲示板など SNS にはユーザの年齢や性別が記載されていない場合が多いため、従来のアンケートのようなプロフィールごとに意見を抽出することはできなかった。(表 1)

これらの課題を解決するため、著者らは SNS 上の投稿内容から投稿者のプロフィールを自動的に推定する手法を考案した。提案手法では、投稿者の過去の投稿内容を遡って取得し、特定のプロフィールに特徴的に現れるキーワードを検出することで、プロフィールを推定する。SNS としては、最近の利用者数の増加傾向、投稿内容の収集や解析に対する規制の少なさなどから、本技術を適用するプラットフォームとして最適と考えられる Twitter を利用した。

以降、2 章において関連研究を紹介、3 章では、提案手法を適用したアプリケーションの利用者となりうる企業に対してヒアリングを実施した結果を紹介し、本稿で扱う問題の性質を述べ、4 章において提案手法の詳細について説明し、5 章では、提案手法の一般的な性能評価実験に加えて、マーケティングにおける前述の 2 つの需要、(1) 投稿者の

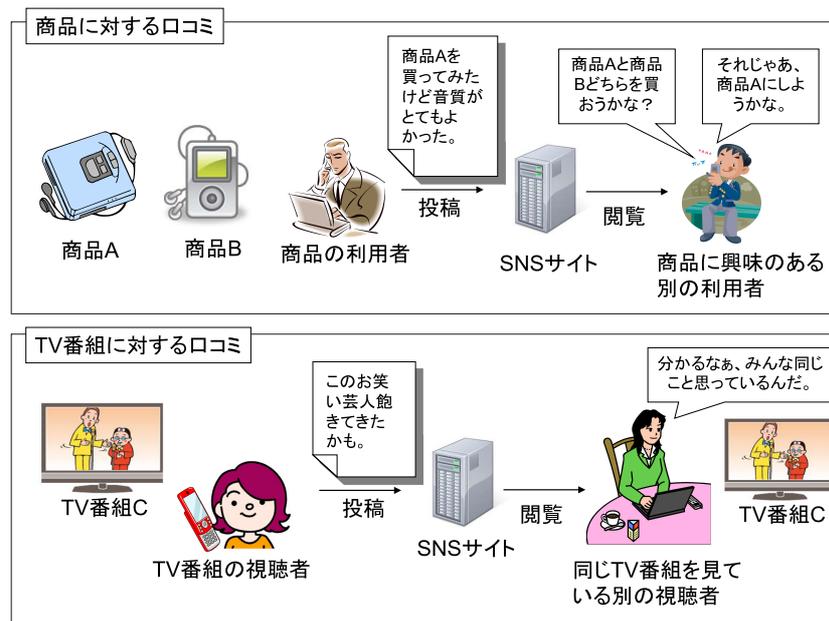


図 1 SNS における口コミ情報の共有

Fig. 1 Social communications related to products and broadcast TV programs.

表 1 SNS 上の口コミ分析とモニタ調査によるアンケートの特性

Table 1 Characteristics of online opinion analysis and questionnaire surveys.

|        | 口コミ分析        | アンケート     |
|--------|--------------|-----------|
| ボリューム  | ○            | △(コストに依存) |
| リアルタイム | ○            | ×         |
| コスト    | ○            | ×         |
| プロフィール | ×            | ○         |
| 正確性    | △(高精度な手法も存在) | ○         |

表 2 プロフィールに対する各社の需要 (○ は必須としてあげた項目, △ はあれば望ましいとしてあげた項目)

Table 2 Demand for demographic types (○ for essential, △ for preferred).

|             | 年齢 | 性別 | 居住地 | 趣味 | 職業 | 年収 | 既未婚 |
|-------------|----|----|-----|----|----|----|-----|
| A 社(商品開発部門) | ○  | ○  | ○   | △  | △  |    |     |
| A 社(宣伝部門)   | ○  | ○  | ○   |    | △  | △  |     |
| B 社         | ○  | ○  | ○   | △  | △  | △  | △   |
| C 社         | ○  | ○  | ○   | △  |    |    |     |
| D 社         | ○  | ○  | ○   | △  | △  |    |     |

分布の推定, (2) 特定のプロフィールを持つ投稿者の検出, についても精度評価を実施した. 6 章では, 提案手法を利用したアプリケーションについて, 動作の仕組みや画面イメージを紹介する.

## 2. 関連研究

Web 上の情報から著者に関する情報を取得する研究はさまざまな観点から行われている. 文献 [5] では, Web ページに記載されている情報が信頼できるものであるかを検出するために, 発信者に関する情報を収集する手法が提案されている. 文献 [6] では, 文章の著者属性を機械学習により分類する研究を次の 3 種類に分類している. (1) 不特定多数の文章に対して, 著者に関する情報をなるべく多く付与する問題 [7], [8], [9], (2) 少数の文章に対して多数の著者候補から著者を同定する問題 [10], (3) ある文章に対して特定の候補者が書いた文章かどうかを判定する問題 [11]. 本稿の課題は (1) に分類される問題と考えられる.

文献 [7], [8], [9] はともに著者の記載した文章に出現する各単語の出現頻度を特徴として識別器を用いてプロフィールを推定する手法を提案している. 文献 [7] では, ブログや学生を対象として収集したエッセイを対象として, 著者の年齢や性別, 母国語, 性格を推定, 文献 [8] では, 英語の E-mail を対象として年齢や性別, 出身地, 学歴, 母国語を推定, 文献 [9] では, ベトナム語のブログを対象として年齢や性別, 居住地を推定している. しかし, これらの研究では, ブログやエッセイ, E-mail など, 独自に用意した実験環境での有効性評価にとどまっており, 実際のアプリケーションへの適用の観点では, 適用方法の提示や有効性の評価には至ってはいない.

本稿では, ネット上の口コミ情報をマーケット分析に応用するためのアプリケーションを構築する課題を解決するために, 多数の 口コミ 情報が収集可能な大規模 SNS プラットフォームである Twitter の投稿者に対してプロフィール推定手法を適用するためのデータ収集方法の提案, マーケット分析を考慮したプロフィール分布精度の評価と再現率, 適合率の評価による有効性の検証, リアルタイムに分析が可能なアプリケーションの実装, を行うことで一連のワークフローを提案する.

## 3. 問題の性質

マーケティングに有効なアプリケーションを提供するため, 異なる業種の企業 4 社の計 5 部門\*1) に対して, 分析を希望するユーザプロフィール情報, 粒度, 活用方法, 要求精度についてヒアリングを実施した.

### 3.1 プロフィールの種類

推定対象とすべきプロフィールを決定するため, 各企業から推定対象としたいプロフィール情報のヒアリングを実施した. 表 2 に示すように, 年齢, 性別, 居住地に関してはどの企業も必須な情報としてあげた. 加えて, 趣味や職業, 年収, 既未婚といった情報も取得できれば望ましいとする企業が見られた. 本稿では, 必須項目としてあげられた年齢, 性別, 居住地を推定する問題を扱うものとし, 他のプロフィールについて同様のアプローチを適用した場合の性能評価は今後の課題とする.

### 3.2 プロフィールの粒度

推定対象とする年齢, 性別, 居住地について, どの粒度で扱うかについてヒアリングを実施した. 年齢については, 4 社ともマーケティングで一般的に利用される, 年代 (10 代, 20 代, 30 代, 40 代など) による分類を扱うとした. B 社からは 50 代以上の情報も取得したいとの意見があったが, 50 代以上の Twitter 利用者はきわめて少なく, 分析結果の有効な活用が困難と想定されたことから, 40 代以上と分類することで合意した. また, 居住地については関東や近畿などの粒度で分かれば十分との意見が多かったため, 日本を 7 地域 (北海道・東北, 関東, 北信越, 東海, 近畿, 中国・四国, 九州・沖縄) に分類して推定することとした.

### 3.3 分析データの活用方法

分析データの活用方法については, 4 社ともそれぞれ指定の商品や話題に対して言及したユーザ集合のプロフィールの分布を知りたいことをあげ, 加えて A 社宣伝部門と B 社

\*1) ヒアリングを実施した企業について, A 社: コンシューマ向け商品, サービスの販売. B 社: インターネットサービスプロバイダ, C 社: メディア, 広告関連事業者, D 社: ビジネス向けネット評判分析ソリューションの販売.

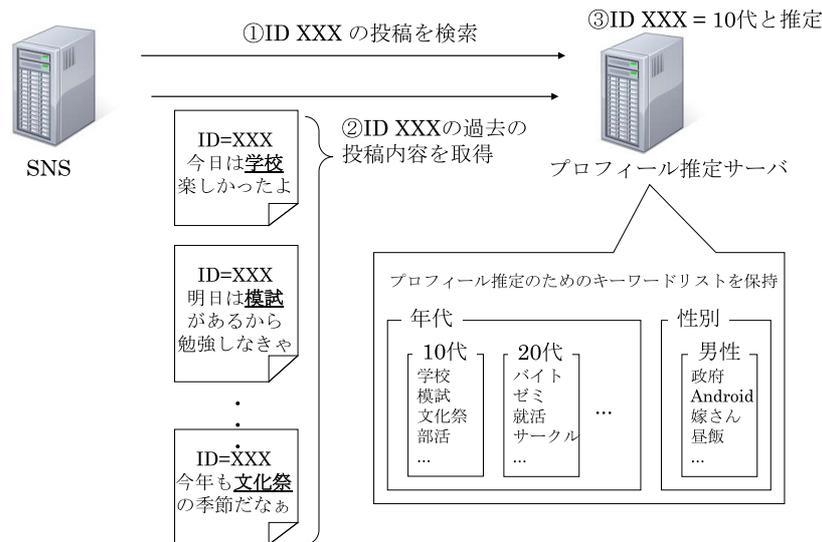


図 2 プロフィール推定技術の動作の仕組み

Fig. 2 Overview of the demographic estimation algorithm.

では、特定の属性を持つユーザに対して Twitter 上で情報発信を行う、といったターゲット広告の目的で本技術を利用することを希望した。

### 3.4 要求分析精度

各企業が求める推定精度は、分析結果をどの程度業務に反映させるかや、業務内容、企業方針などによって多様であると考えられ、また分析対象とする個々の商品や話題によっても異なると考えられるため、一概に定義することはできないが、上記 4 社へのヒアリングにおいては、前述の 2 種類の活用方法において、それぞれ 90%以上の精度があれば、分析結果を信頼して活用できるというものが、最も高い要求精度であった。

## 4. プロフィール推定技術

### 4.1 技術の概要

図 2 に本技術の概要を示す。本技術では、投稿に付与された投稿者 ID (アカウント名やユーザ名など) を利用して、投稿者の過去の複数のコメントを取得し、その中に含まれているプロフィール推定に役立つキーワードを検出することでプロフィールを推定する。たとえば、日常のコメントに「学校」や「模試」、「部活」などのキーワードが頻繁に見られる投稿者は、年齢が「10代」で職業が「学生」などと推定することが可能である。

本技術では、Twitter 投稿者のプロフィール欄に記載されたプロフィール情報を人手によって分類し、それらを用いてプロフィール推定のためのキーワードリストを、統計的指標である AIC (赤池情報量基準) [12] に基づいて自動的に構築する。さらに、人工知能の分野で利用される識別器 SVM (Support Vector Machine) [13] を用いて、プロフィールごとのキーワードの出現傾向を学習することで、

高精度な推定を実現した。4.2 節では、プロフィール情報を効果的に収集するための方法、4.3 節では、プロフィール推定のためのキーワードリスト生成方法、4.4 節では、SVM を利用したプロフィール推定方法をそれぞれ記述する。

### 4.2 プロフィール情報の収集

プロフィールを自動的に推定するためのキーワードリストの構築と、識別器の学習の両方の目的で利用する、プロフィールが既知であるユーザの情報を収集する方法について説明する。

Twitter では、投稿者のプロフィールとして、自由文形式で自己紹介文を記載可能なフィールドと、位置情報を記載するフィールドが存在し、それぞれ年代や性別の情報や都道府県名を記載しているユーザが少数であるが存在するため、これらの収集を行う。収集手順として、投稿された Tweet を Twitter 社が提供する streaming API を用いてランダムに収集し、Tweet 本文中に日本語の文字コードを 1 文字以上含むユーザについて、プロフィール情報を取得する。

プロフィール情報中の記載において、年代については、「37 歳」のように直接的に年齢を特定できる場合、「1974 年生まれ」のように生年から年齢を特定できる場合、「30 代」のように年代を記載している場合、などがある。性別については、「男」、「男子」、「オヤジ」など男性を表す表現を含む場合、「女」、「女子」、「おばさん」など女性を表す表現を含む場合、などがある。これらの表現を表記の揺らぎを考慮した正規表現 (算用数字や漢数字、カタカナ表記など) によって、収集したプロフィール情報に対して、検索することで絞り込みを行う。このとき、「15 歳の息子を持つ母」は 15 歳ではないように、必ずしも自動検索が正しいとは限らないため、絞り込みを行ったプロフィール情報

表 3  $E(s)$  値算出に用いる文字列  $s$  の出現回数  
Table 3 Criteria for calculation of  $E(s)$ .

|            |             |              |       |
|------------|-------------|--------------|-------|
|            | 文字列 $s$ が出現 | 文字列 $s$ が非出現 | 合計    |
| $D_p$ の文書数 | $N_{11}(s)$ | $N_{12}(s)$  | $N_p$ |
| $D_n$ の文書数 | $N_{21}(s)$ | $N_{22}(s)$  | $N_n$ |
| 合計         | $N(s)$      | $N(\neg s)$  | $N$   |

を人手で確認する。確認作業は正確性を期すため、2人の作業者が独立して確認を行い、両者が正しいと判定したプロフィール情報のみを利用するものとする。

また、絞り込みに用いるキーワードについては、「高校生」などは10代、「サラリーマン」などは男性を間接的に表すキーワードととらえて利用することで、より大規模なデータを収集することができるが、収集されるユーザが特定のプロフィールに偏った分布になることを避けるため、本稿では利用しないものとする。

### 4.3 キーワードリストの生成

プロフィール推定に役立つキーワードリストは、特定のプロフィールを持つ（たとえば年代が10代である、など）投稿者の投稿内容に偏って現れる文字列を、統計的な指標を用いて検出することで構築する。

学習用文書のうち、特定のプロフィールを持つ投稿者が投稿した文書を  $D_p$ 、それ以外のプロフィールを持つ投稿者が投稿した文書を  $D_n$  とする。ある文字列  $s$  が  $D_p$  に偏って出現する度合いを表す指標  $E(s)$  を、著名な統計指標である AIC（赤池情報量基準）[12] を用いて算出する。表 3 のように、ある文字列  $s$  が出現する  $D_p$  の文書数  $N_{11}$  と  $D_n$  の文書数  $N_{21}$ 、文字列  $s$  が出現しない  $D_p$  の文書数  $N_{12}$  と  $D_n$  の文書数  $N_{22}$  の4つの値を、学習用文書に出現するすべての文字列について求め、 $E(s)$  値の大きい上位2,000件をプロフィール推定に役立つキーワードとする。

文字列  $s$  が  $D_p$  に偏って出現する度合い  $E(s)$  は文献 [14] の知見をもとに、AIC の独立モデルに対する値  $AIC_{IM}$  および従属モデルに対する値  $AIC_{DM}$  を用いて、次のように定義することができる。

$$N_{11}(s)/N(s) > N_{12}(s)/N(\neg s) \text{ のとき,}$$

$$E(s) = AIC_{IM}(s) - AIC_{DM}(s) \tag{1}$$

$$N_{11}(s)/N(s) \leq N_{12}(s)/N(\neg s) \text{ のとき,}$$

$$E(s) = AIC_{DM}(s) - AIC_{IM}(s)$$

ここで、 $AIC_{IM}(s)$ 、 $AIC_{DM}(s)$  はそれぞれ文献 [12] の定義に従って、次の式で与えられる。

$$AIC_{IM}(s) = -2 \times MLL_{IM} + 2 \times 2$$

$$MLL_{IM} = N_p(s) \log N_p(s) + N(s) \log N(s)$$

$$+ N_n(s) \log N_n(s) + N(\neg s) \log N(\neg s) - 2N \log N$$

$$AIC_{DM}(s) = -2 \times MLL_{DM} + 2 \times 3$$

表 4 文字列の出現回数と  $E(s)$  値の例

Table 4 Examples of term frequencies and  $E(s)$  values.

| 文字列 | $N_{11}(s)$ | $N_{12}(s)$ | $N_{21}(s)$ | $N_{22}(s)$ | $E(s)$ |
|-----|-------------|-------------|-------------|-------------|--------|
| 模試  | 261         | 61          | 639         | 2640        | 2640   |
| ビール | 70          | 819         | 830         | 1882        | -216.6 |
| バケツ | 17          | 51          | 883         | 2650        | -2.0   |

$$MLL_{DM} = N_{11}(s) \log N_{11}(s) + N_{12}(s) \log N_{12}(s)$$

$$+ N_{21}(s) \log N_{21}(s) + N_{22}(s) \log N_{22}(s) - N \log N \tag{2}$$

具体例として、10代の投稿者が記載した文書に偏って出現する文字列「模試」と10代でない投稿者が記載した文書に偏って出現する文字列「ビール」、偏りなく出現する文字列「バケツ」の出現回数と  $E(s)$  値の例を表 4 に示す。「模試」は10代に偏って出現する文字列であるため、10代である度合いを表す指標  $E(s)$  が正の値をとり、「ビール」は10代以外に偏っているため  $E(s)$  は負の値をとる。「バケツ」は偏りなく出現するため、 $E(s)$  は0に近い値となる。

### 4.4 SVM (Support Vector Machine) を利用したプロフィール推定

プロフィールの推定には人工知能分野で用いられる識別器 SVM を用いる。SVM を用いて学習用文書におけるキーワードの出現傾向を学習することで、プロフィールが未知である投稿者に対して、その投稿内容からプロフィールを推定することが可能となる。具体的には、プロフィールごとに抽出したキーワード  $S_1, S_2, S_3, \dots, S_n$  と、学習用文書それぞれにおける各キーワードの出現回数  $M_1, M_2, M_3, \dots, M_n$  からなる行列を SVM の入力として与える。学習段階では、当該プロフィールであるか、当該プロフィール以外であるかを表すラベル Label もあわせて与えることで SVM を学習させる。表 5 に SVM の入力例を示す。

利用するキーワード数  $n$  は学習用文書に出現するすべての単語数とすることで最も性能が高くなるが、キーワード数が多いと計算時間が大きくなることが課題である。この問題を解決するため、4.2 節で説明したような統計的指標を用いて有効性の高いキーワードのみに絞り込むことで  $n = 2000$  程度に抑制し、高速・高精度なプロフィール推定を実現した。

表 5 SVM の入力となる特徴量の例  
Table 5 Example of input of SVMs.

|       | $S_1$    | $S_2$    | $S_3$    | ... | $S_n$    | Label<br>(学習データのみ) |
|-------|----------|----------|----------|-----|----------|--------------------|
| 投稿者 1 | $M_{11}$ | $M_{12}$ | $M_{13}$ | ... | $M_{1n}$ | 1                  |
| 投稿者 2 | $M_{21}$ | $M_{22}$ | $M_{23}$ | ... | $M_{2n}$ | 0                  |
| ...   | ...      | ...      | ...      | ... | ...      | ...                |
| 投稿者 x | $M_{x1}$ | $M_{x2}$ | $M_{x3}$ | ... | $M_{xn}$ | 0                  |

表 6 収集した Twitter 投稿者の年齢ごとの人数と比率  
Table 6 Characteristics of collected user ages.

| 端数(歳) | 10代  |      | 20代  |      | 30代  |      | 40代以上 |      |
|-------|------|------|------|------|------|------|-------|------|
|       | 人    | %    | 人    | %    | 人    | %    | 人     | %    |
| 0     | 4    | 0.4  | 306  | 8.4  | 160  | 10.2 | 80    | 7.8  |
| 1     | 4    | 0.4  | 228  | 6.3  | 131  | 8.3  | 85    | 8.3  |
| 2     | 4    | 0.4  | 354  | 9.7  | 118  | 7.5  | 60    | 5.9  |
| 3     | 18   | 1.6  | 320  | 8.8  | 114  | 7.2  | 62    | 6.1  |
| 4     | 18   | 1.6  | 330  | 9.1  | 88   | 5.6  | 27    | 2.6  |
| 5     | 57   | 5.1  | 332  | 9.1  | 104  | 6.6  | 49    | 4.8  |
| 6     | 123  | 11.1 | 306  | 8.4  | 77   | 4.9  | 31    | 3.0  |
| 7     | 167  | 15.1 | 337  | 9.3  | 67   | 4.3  | 45    | 4.4  |
| 8     | 326  | 29.4 | 209  | 5.7  | 54   | 3.4  | 24    | 2.4  |
| 9     | 300  | 27.1 | 187  | 5.1  | 46   | 2.9  | 27    | 2.6  |
| 年代    | 88   | 7.9  | 726  | 20.0 | 615  | 39.1 | 408   | 40.0 |
| 50代以上 |      |      |      |      |      |      | 122   | 12.0 |
| 合計    | 1109 |      | 3635 |      | 1574 |      | 1020  |      |

表 7 実験に利用したプロフィールごとの投稿者数 (人)  
Table 7 Number of users in the experiment.

| 年代    | 投稿者数  | 性別 | 投稿者数  | 居住地域   | 投稿者数  |
|-------|-------|----|-------|--------|-------|
| 10代   | 1,000 | 男性 | 2,560 | 北海道・東北 | 992   |
| 20代   | 1,000 | 女性 | 2,560 | 関東     | 992   |
| 30代   | 1,000 |    |       | 北信越    | 992   |
| 40代以上 | 1,000 |    |       | 東海     | 992   |
|       |       |    |       | 近畿     | 992   |
|       |       |    |       | 中国・四国  | 992   |
|       |       |    |       | 九州・沖縄  | 992   |
| 計     | 4,000 | 計  | 5,120 | 計      | 6,944 |

## 5. 性能評価実験

### 5.1 実験環境と実験手順

**実験環境:** 計算機 (1core 3.2GHz, 8GB RAM, Cent OS), 提案手法で利用する SVM として Lib SVM [15], 文書から単語を文字列として切り出すために形態素解析器 MeCab [16] を用いた. 実装には C 言語を用いた.

**利用データ:** 4.2 節に記載の手順により, Twitter のプロフィール情報 10 万人分を収集し, 年代, 性別, 居住地域のいずれかについて記載のある投稿者を人手によってプロフィールごとに分類した. 具体例として, 年代については表 6 に示す構成比率でユーザを収集した. 若年層と年配層

の Twitter 投稿者は少ないものの, それ以外は偏りなく収集できている. この中で最も少ない 40 代に合わせて, 各年代から 1,000 件ずつを構成比に合わせて取得し, 実験に利用した. 性別, 居住地域についても同様の方法で, 表 7 に示す投稿者数を収集した. 各プロフィールを持つ投稿者数の半数を本技術の学習のために利用し, 残りの半数のプロフィールを推定した. 学習に利用していない投稿者に対する推定精度は, 実運用時の推定精度と同じと考えることができる.

**実験手順:** 提案手法を用いて各プロフィールに特徴的に現れるキーワードを抽出する. 次に, 投稿者の投稿内容から本技術を用いてプロフィール (年代, 性別, 居住地域)

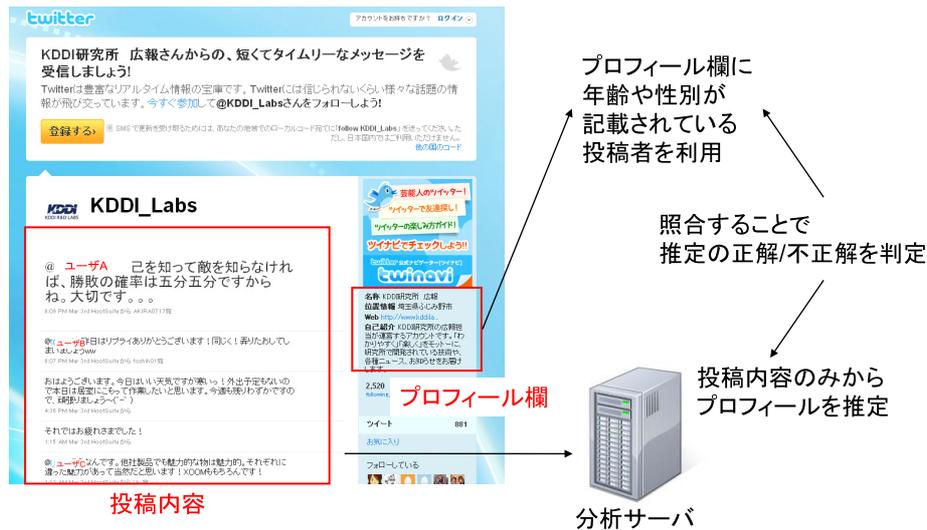


図 3 プロフィール推定精度の評価方法

Fig. 3 Evaluation of the accuracy of demographic estimation.

を推定，実際のプロフィールと照合して正解を判定する (図 3)．推定精度については，次の 3 つの方法で評価した．(1) 手法の一般的な性能を評価するため，正解率 (正しく推定した投稿者数/特定のプロフィールを持つ投稿者数) を評価，分母は 10 代であれば，本稿では 500 人となる．(2) ある商品や話題に対して言及した Twitter ユーザの分布 (年代であれば，10 代，20 代，30 代，40 代以上が各何ユーザで構成されているか) を把握したいといったニーズを想定し，推定対象となるユーザ集合中の実際の分布と，推定によって得られた分布の誤差の評価を，推定対象の分布が均等，不均等それぞれの場合について実施，推定されたプロフィールの分布誤差率  $E$  をプロフィールごとに評価し，さらにプロフィールの種類 (年代，性別，居住地) ごとの平均誤差率  $E_{avg}$  を評価した．誤差率  $E$ ，平均誤差  $E_{avg}$  はプロフィールの種類を  $T$ ， $T$  の各要素  $T_1, T_2, \dots, T_n$  ( $T =$  年代であれば， $n = 4$ ， $T_1 = 10$  代， $T_2 = 20$  代， $T_3 = 30$  代， $T_4 = 40$  代以上)， $T_i$  の実際の投稿者数  $U_t(T_i)$ ， $T_i$  と推定された投稿者数  $U_e(T_i)$  を用いて，それぞれ次のように定義する．

$$E(T_i) = U_e(T_i)/U_t(T_i) - 1$$

$$E_{avg}(T) = \frac{\sum_1^n (|U_t(T_i) - U_e(T_i)|)}{\sum_1^n U_t(T_i)} \quad (3)$$

(3) ターゲット広告や特定のユーザ層を対象としたマーケティングを想定し，特定のプロフィールを持つ投稿者の検出精度を再現率 (正しく推定した投稿者数/特定のプロフィールを持つ投稿者数) と適合率 (正しく推定した投稿者数/特定のプロフィールを持つと推定された投稿者数) の関係で評価した．適合率は誤判定の少なさ，再現率は検出漏れの少なさを示し，すべての推定対象ユーザに対して推定結果を付与した際の再現率は前述の正解率と等しくなる

ものとする．具体例として，提案手法によって 200 人が 10 代であると推定され，その中で本当に 10 代であるユーザ数が 100 人であったとすると，再現率 =  $100/500 = 20\%$ ，適合率 =  $100/200 = 50\%$ となる．再現率と適合率のトレードオフの評価は，SVM 内部で算出される推定の信頼度が高いユーザを優先的に判定することにより行う．具体的には，年代であれば各ユーザに対して 10 代，20 代，30 代，40 代以上のそれぞれに対して推定信頼度を算出し，それらの総和が 1 となるよう正規化を行い，信頼度が最大である年代を当該ユーザの年代として推定する．このとき，信頼度が高いユーザほど，推定結果が確からしいと考えられるため，推定信頼度が高い順に上位 10%，20%...100%と判定することで，再現率と適合率のトレードオフを算出することが可能となる．

### 5.2 実験結果

表 8 に提案手法によって自動的に抽出されたプロフィールを推定するためのキーワードリストの一部，表 9 に (1) のプロフィールごとの推定精度，表 10，表 11，表 12 に (2) の推定対象の分布に対する推定された投稿者分布の比較，表 13 に (3) の特定のプロフィールを持つ投稿者の検出における再現率，適合率の関係の一部を示す．

表 8 は年代や性別，居住地域について，上記の方法で抽出したキーワードリストの  $E(s)$  値が高い上位の一部を例示している．年代については，10 代は学校生活に関するキーワードが検出された．20 代は大学や就職活動に関するキーワードが上位として検出された．表の範囲にはないが，「新入社員」や「初任給」など企業に関するものも見られた．30 代は仕事や家庭に関するキーワードが多いことが分かった．40 代は家庭や政治，自身の健康に対するキーワードが多く見られた．性別では，男女ともに配偶者を指

表 8 抽出されたキーワードリストの一例 (地域は国内を 7 地域に分類したうちの 2 地域について掲載)

Table 8 Example of extracted terms (two of seven areas are shown as samples).

| 10 代 | 20 代 | 30 代 | 40 代以上 | 男性      | 女性   | 関東   | 近畿    |
|------|------|------|--------|---------|------|------|-------|
| 数学   | 大学   | 仕事   | 息子     | 政府      | 旦那   | 新宿   | 大阪    |
| 学校   | バイト  | 会社   | アラフォー  | Android | 母    | 池袋   | 梅田    |
| 模試   | ゼミ   | 出勤   | ランチ    | 嫁さん     | お風呂  | 渋谷   | 京都    |
| テスト  | 就活   | 職場   | 運動会    | 昼飯      | 洗濯   | 横浜   | やけど   |
| 国語   | 履修   | 上司   | 血压     | 企業      | 夫    | 山手線  | ほんま   |
| 体育   | 講義   | 三十路  | 政権     | 政治      | お母さん | 電車   | 阪神    |
| 文化祭  | サークル | 残業   | 外交     | マスコミ    | ☆    | 新橋   | 神戸    |
| 受験生  | 休講   | 出社   | 家内     | 奥さん     | 友達   | 吉祥寺  | 心齋橋   |
| 部活   | 単位   | ビール  | 休肝日    | 国民      | ご飯   | 秋葉原  | マクド   |
| 受験   | 学祭   | 独身   | ゴルフ    | Google  | お弁当  | 京浜東北 | サンテレビ |

表 9 プロフィールごとの推定精度

Table 9 Estimation accuracy of the proposed algorithm.

| 年代     | 正解率(%) | 性別 | 正解率(%) | 居住地域   | 正解率(%) |
|--------|--------|----|--------|--------|--------|
| 10 代   | 83.8   | 男性 | 90.7   | 北海道・東北 | 77.2   |
| 20 代   | 68.0   | 女性 | 86.1   | 関東     | 61.1   |
| 30 代   | 50.7   |    |        | 北信越    | 69.8   |
| 40 代以上 | 69.4   |    |        | 東海     | 72.8   |
|        |        |    |        | 近畿     | 72.0   |
|        |        |    |        | 中国・四国  | 65.7   |
|        |        |    |        | 九州・沖縄  | 76.2   |
| 全体     | 68.0   | 全体 | 88.0   | 全体     | 70.8   |

表 10 推定対象の分布 (均等) と推定された分布の比較

Table 10 Distribution error of each demographic (for uniform distribution).

| プロフィール | 推定対象の分布<br>$U_i(T_i)$ (均等) | 推定された分布<br>$U_e(T_i)$ | 誤差率(%)<br>$E(T_i)$ |
|--------|----------------------------|-----------------------|--------------------|
| 10 代   | 500                        | 500                   | 0.0                |
| 20 代   | 500                        | 440                   | -12.0              |
| 30 代   | 500                        | 460                   | -8.0               |
| 40 代以上 | 500                        | 600                   | 20.0               |
| 男性     | 1,280                      | 1,349                 | 5.4                |
| 女性     | 1,280                      | 1,211                 | -5.4               |
| 北海道・東北 | 496                        | 520                   | 4.8                |
| 関東     | 496                        | 538                   | 8.5                |
| 北信越    | 496                        | 488                   | -1.6               |
| 東海     | 496                        | 487                   | -1.8               |
| 近畿     | 496                        | 510                   | 2.8                |
| 中国・四国  | 496                        | 440                   | -11.3              |
| 九州・沖縄  | 496                        | 489                   | -1.4               |

す言葉は特徴的といえる。男性は仕事や政治, IT などへの興味が高く, 女性は家事や食べ物への興味が高いことが分かった。居住地域では, 地名や地域固有の交通機関, テレビ局名, 方言などが見られた。特に Twitter では, 現在自分がどこで何をしているかを発信するため, 地名に関する情報は多く得られると考えられる。

表 9 から年代については 10 代の推定精度が高く, 30 代の推定精度が低いことが分かった。10 代の多くは学生と考えられ, 類似した生活傾向を持つために投稿内容に類似性が現れ, 高精度な推定が可能であったと考えられるのに対し, 30 代は 20 代後半および 40 代前半と類似した生活傾向を持つと考えられ, それらの区別が困難であったため

表 11 推定対象の分布（不均等）と推定された分布の比較

Table 11 Distribution error of each demographic (for biased distribution).

| プロフィール | 推定対象の分布 $U_i(T_i)$<br>(不均等) | 推定された分布<br>$U_e(T_i)$ | 誤差率 (%)<br>$E(T_i)$ |
|--------|-----------------------------|-----------------------|---------------------|
| 10代    | 400                         | 372                   | -7.0                |
| 20代    | 300                         | 282                   | -6.0                |
| 30代    | 200                         | 184                   | -8.0                |
| 40代以上  | 100                         | 162                   | 62.0                |
| 男性     | 1,280                       | 1,210                 | -5.5                |
| 女性     | 320                         | 390                   | 21.9                |
| 北海道・東北 | 150                         | 180                   | 20.0                |
| 関東     | 450                         | 390                   | -13.3               |
| 北信越    | 150                         | 159                   | 6.0                 |
| 東海     | 150                         | 171                   | 14.0                |
| 近畿     | 300                         | 255                   | -15.0               |
| 中国・四国  | 150                         | 150                   | 0.0                 |
| 九州・沖縄  | 150                         | 195                   | 30.0                |

表 12 プロフィールの種類ごとの分布誤差平均  $E_{avg}$

Table 12 Average distribution error of each demographic.

|     | 誤差平均 $E_{avg}$ (均等) | 誤差平均 $E_{avg}$ (不均等) |
|-----|---------------------|----------------------|
| 年代  | 10.0                | 12.4                 |
| 性別  | 5.4                 | 8.8                  |
| 居住地 | 4.6                 | 14.0                 |

に、推定精度が低下したと考えられる。性別については、男性、女性ともに高い推定精度となった。居住地域については関東の推定精度が低かった。これは、他の地域の投稿者も首都圏について言及することが多く、区別が困難なためと考えられる。

表 10, 11 の実験では、年代、性別、居住地それぞれについて推定対象の分布を表に記載の値を設定し、分布誤差を評価した。分布の均等、不均等ともに、年代、性別、居住地ともに推定対象の分布に近い推定結果が得られていることが分かる。表 9 の推定精度に比べて分布の誤差が小さいのは、表 9 では投稿者ごとに正解・不正解を判定しているが、たとえば 20代と 30代を取り違えて判定したとしても誤りが相殺されて分布の推定結果の誤差は小さくなるためと考えられる。表 12 から、分布が均等に近いほうが誤差は少なくなる傾向が得られた。これは均等な分布のデータを SVM の学習に利用したことが原因と考えられる。プロフィールの分布を推定するために最適な学習データの作成方法については今後の課題であるが、本稿では汎用的な手法として均等分布のデータを学習に用いた。

表 13 は検出する推定信頼度を上位 10%, 20%, 30%... 100%まで 10%刻みで変化させたときのプロフィールごとの再現率と適合率の関係を示している（紙面の都合上、適合率が高い部分を抜粋して掲載）。年代については、特に 10代、20代を高適合率で検出することが可能であること

が分かった。10代については、再現率 25.1%（10代の全ユーザの 25.1%）を検出したときの適合率は 100%（推定結果はすべて正解）であった。性別や居住地についても、再現率が低い領域での適合率はきわめて大きい。この技術を利用すれば、特定のユーザ層（たとえば 10代女性など）の意見を商品やテレビ番組の企画に反映させたいような利用シーンにおいて、当該プロフィールを持つユーザの意見のみを収集して、内容の閲覧、分析を行うことが可能となる。

最後に、本技術を用いたプロフィール推定に要する処理時間を評価したところ、年代、性別、居住地域のすべてを推定しても 1投稿者あたり平均 1秒以下であった。国内の Twitter ユーザが 1200 万ユーザであるとする、1台の PC のみを用いた場合でも、約 4 カ月程度で全ユーザの推定が完了する計算になる。また、テレビ番組に対する意見をリアルタイムに分析するような利用シーンにおいて、アプリケーションでまだ推定されていないユーザが出現したとしても、一定量（たとえば 200 件）以上の Tweet があれば、リアルタイムにプロフィールを推定することが可能である。

### 5.3 考察および改善案

3章で設定したプロフィールの分布誤差、再現率と適合率のトレードオフの観点から本稿での精度を考察すると、分布に偏りのあるデータを推定対象とした場合に、年代や居住地で推定誤差が 10%以上となる点は性能改善が必要と考えられる。また、適合率 90%以上における再現率を検証すると、20代、30代、40代以上および関東に対する推定精度は向上が必要であるといえる。

年代の推定においては、20代後半と 30代前半、30代後半と 40代前半の生活傾向が類似している点が性能低下の原因として考えられる。これらは分類粒度を詳細化（たと

表 13 プロフィールごとの再現率  $R$  (%), 適合率  $P$  (%) の関係

Table 13 Comparison of the recall (%) and the precision (%) of the proposed algorithm.

| 信頼度    | 10代  |       | 20代  |      | 30代  |      | 40代  |      |
|--------|------|-------|------|------|------|------|------|------|
|        | R    | P     | R    | P    | R    | P    | R    | P    |
| 上位 10% | 25.1 | 100.0 | 4.2  | 95.0 | 2.2  | 50.0 | 4.0  | 66.7 |
| 20%    | 40.0 | 98.9  | 13.1 | 93.7 | 6.4  | 63.0 | 10.9 | 71.0 |
| 30%    | 49.1 | 97.4  | 21.8 | 91.6 | 11.6 | 64.2 | 19.3 | 69.6 |
| 40%    | 54.9 | 95.4  | 28.7 | 89.6 | 16.0 | 55.4 | 28.8 | 69.5 |
|        | ...  | ...   | ...  | ...  | ...  | ...  | ...  | ...  |
| 100%   | 83.8 | 83.8  | 68.0 | 77.3 | 50.7 | 55.1 | 69.4 | 57.9 |

| 信頼度    | 男性   |      | 女性   |      | 北海道・東北 |       | 関東   |       |
|--------|------|------|------|------|--------|-------|------|-------|
|        | R    | P    | R    | P    | R      | P     | R    | P     |
| 上位 10% | 19.6 | 98.0 | 0.0  | 0.0  | 13.7   | 100.0 | 4.4  | 100.0 |
| 20%    | 39.5 | 98.6 | 0.0  | 0.0  | 27.0   | 97.8  | 8.1  | 97.6  |
| 30%    | 52.7 | 97.1 | 5.7  | 98.6 | 37.1   | 97.4  | 14.7 | 98.6  |
| 40%    | 58.8 | 96.9 | 19.2 | 99.2 | 44.2   | 94.8  | 21.0 | 93.7  |
|        | ...  | ...  | ...  | ...  | ...    | ...   | ...  | ...   |
| 100%   | 90.7 | 86.1 | 86.1 | 90.2 | 77.2   | 73.7  | 61.1 | 56.3  |

| 信頼度    | 北信越  |      | 東海   |      | 近畿   |      | 中国・四国 |      | 九州・沖縄 |      |
|--------|------|------|------|------|------|------|-------|------|-------|------|
|        | R    | P    | R    | P    | R    | P    | R     | P    | R     | P    |
| 上位 10% | 8.3  | 97.6 | 10.5 | 96.3 | 5.2  | 92.9 | 8.1   | 97.6 | 18.1  | 97.8 |
| 20%    | 19.0 | 96.9 | 21.6 | 96.4 | 13.9 | 97.2 | 17.7  | 98.9 | 29.2  | 98.0 |
| 30%    | 28.6 | 95.9 | 30.6 | 91.6 | 24.4 | 94.5 | 25.4  | 96.9 | 39.1  | 94.2 |
| 40%    | 36.7 | 94.8 | 39.3 | 88.6 | 35.7 | 88.9 | 33.9  | 94.9 | 47.4  | 91.1 |
|        | ...  | ...  | ...  | ...  | ...  | ...  | ...   | ...  | ...   | ...  |
| 100%   | 69.8 | 70.9 | 72.8 | 74.1 | 72.0 | 70.0 | 65.7  | 74.1 | 76.2  | 77.3 |

えば5歳区切り)にすることで、SVMによる識別性能を高め、高精度化の可能性が見込める。

本稿では、収集したユーザプロフィールをそのまま学習に利用した場合の性能評価を実施したが、小学生や50代以上の年配層は学習データが不足しているため、精度は十分ではないと考えられる。具体的には、10代の実験結果において年齢が15歳以下と記載した47名のユーザについて正解率を評価したところ68.1%であり、10代全体の正解率83.8%と比較して低いものであった。小学生や50代以上の年配層などは、Twitter投稿者数がきわめて少なく、本稿で提案するアプリケーションで有効な活用が困難と考えられるため、これらを学習データから除くことで、実用性を高められる可能性がある。

## 6. アプリケーションへの応用

本技術を応用した、商品やコンテンツに対するネット上の意見を分析して、マーケティングなどへ利用するための口コミ分析アプリケーションを紹介する。

**利用シーン:** 本アプリは企業における新商品の企画やプロモーション効果測定のためにTwitter上の意見を収集、分析してPC画面などで閲覧するために利用できる。投稿者のプロフィールが取得できることで、ターゲットを限定した評判分析、商品へのニーズ調査などが可能となる



図 4 口コミ分析アプリケーションの利用シーン

Fig. 4 Use case of the application.

(図 4)。

**動作の仕組み:** 商品やコンテンツに対して投稿された意見をサーバで収集し、付与された投稿者IDをもとに提案手法を用いてプロフィールを推定する。Twitterでは特定の議題に対する書き込みを表すタグ(ハッシュタグと呼ばれる)が存在し、テレビ番組に対するハッシュタグを利用して意見を共有するユーザも多いため、ハッシュタグを検索することで意見の収集が可能となる。また、商品については商品名を含む投稿を検索することも有効である。加えて、肯定的/否定的な表現を多数収録した評判解析用の辞書を用いて、意見の内容が肯定的であるか否定的であるかを判定し、これらをあわせてネット意見集約コンテンツとしてユーザの端末に配信する(図5)。

**アプリケーション画面:** アプリケーションの画面イメー

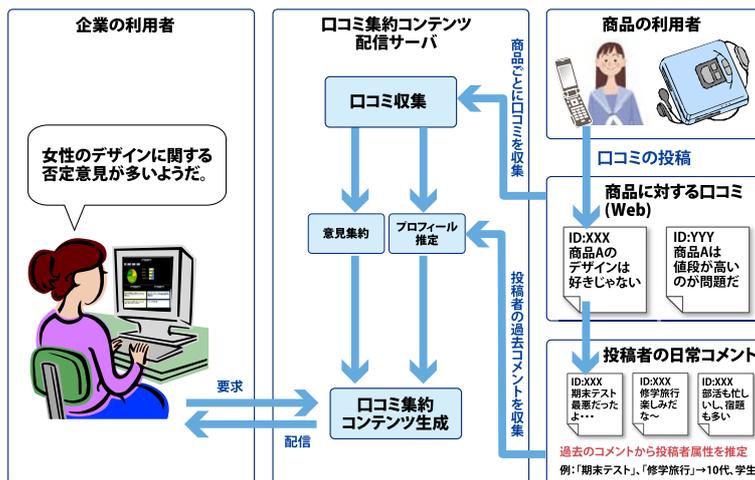


図 5 口コミ分析アプリケーションの動作の仕組み  
Fig. 5 Configuration of the application.



図 6 口コミ分析アプリケーションの画面イメージ  
Fig. 6 Screenshot of the Web interface of the application.

ジを図 6 に示す。画面左には調査対象に設定した商品やテレビ番組のリストが並んでおり、クリックすることで画面右上に投稿者のプロフィールごとの分布と肯定/否定意見の割合が表示される。画面右下には、肯定/否定意見の具体的な内容が表示され、どのような点が肯定され、どのような点が否定されたのかを手軽に確認することができる。肯定/否定意見は年代や性別を絞り込んで表示させることも可能である。

他言語への適用：提案手法を日本語以外の言語に適用する場合、形態素解析器を対象言語に適したものに置換し、本稿に記載の方法で当該言語の Twitter ユーザ情報を収集することで、適用が可能と考えられる。

## 7. まとめ

本稿で紹介したプロフィール推定技術は、従来の評判解

析では取得できなかった投稿者のプロフィールを取得することで、膨大なネット上の意見をリアルタイムに低コストで分類、集約することを可能とした。本技術はマーケティングを中心に、幅広い分野への応用が可能であり、紹介した口コミ分析アプリケーションは、実際のテレビ番組において視聴者のネット上の意見をリアルタイムに集計するサービスとして実用化されている。今後は製造業や食品、サービスなどを対象としたマーケティングなど、さらに多くの分野への展開を進めていく予定である。

## 参考文献

- [1] Twitter, <http://twitter.com/>
- [2] mixi, <http://mixi.jp/>
- [3] IT media: mixi, Twitter, Facebook 2011 年 1 月最新ニールセン調査, 2011 年 2 月 21 日, 入手先 (<http://blogs.itmedia.co.jp/saito/2011/02/>)

mixi-twitter-fa.html).

[4] Dave, K., Lawrence, S. and Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, *WWW*, pp.519-528 (2003).

[5] Kato, Y., Kawahara, D., Inui, K., Kurohashi, S. and Shibata, S.: Identifying the Information Sender Configuration of Web Pages, *Proc. 2009 IEEE/ACM/WIC International Conference on Web Intelligence (WI'09)*, pp.335-340 (2009).

[6] Koppel, M., Schler, J. and Argamon, S.: Computational Methods in Authorship Attribution, *Journal of the American Society for Information Science and Technology*, Vol.60, No.1, pp.9-26 (2009).

[7] Argamon, S., Koppel, M., Pennebaker, J. and Schler, J.: Automatically Profiling the Author of an Anonymous Text, *Comm. ACM*, Vol.52, No.2, pp.119-123 (2009).

[8] Estival, D., Gaustad, T., Pham, S.B., Radford, W. and Hutchinson, B.: Author Profiling for English Emails, *Proc. 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pp.262-272 (2007).

[9] Pham, D.D., Tran, G.B. and Pham, S.B.: Author Profiling for Vietnamese Blogs, *Proc. International Conference on Asian Language Processing*, pp.190-194 (2009).

[10] Abbasi, A. and Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection, *ACM Trans. Inf. Syst.*, Vol.26, Issue 2, No.7 (2008).

[11] Koppel, M., Schler, J. and Bonchek-Dokow, E.: Measuring Differentiability: Unmasking Pseudonymous Authors, *Journal of Machine Learning Research*, Vol.8, pp.1261-1276 (2007).

[12] Akaike, H.: A New Look at the Statistical Model Identification, *IEEE Trans. Automatic Control*, Vol.19, No.6, pp.716-723 (2003).

[13] Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, pp.273-297 (1995).

[14] Matsumoto, K. and Hashimoto, K.: Schema Design for Causal Law Mining from Incomplete Database, *Proc. Discovery Science: 2nd International Conference (DS'99)*, pp.92-102 (1999).

[15] Fan, R., Chen, P. and Lin, C.: Working Set Selection Using Second Order Information for Training SVM, *Journal of Machine Learning Research*, Vol.6, pp.1889-1918 (2005), available from (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>).

[16] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp.230-237 (2004).



池田 和史 (正会員)

2006年大阪大学基礎工学部情報科飛び級のため中退。2008年同大学大学院博士前期課程修了。同年KDDI(株)入社, 研究所所属。自然言語処理, ソーシャルメディア解析等の研究に従事。電子情報通信学会, 日本データベース

学会各会員。



服部 元 (正会員)

1996年神戸大学工学部電気電子工学科卒業。1998年神戸大学大学院自然科学研究科電気電子工学専攻修士課程修了。同年国際電信電話(株)(現KDDI(株))入社。現在, (株)KDDI研究所知能メディアグループ研究員。

この間, ネットワーク管理, ITS, ソフトウェアエージェント, Webコンテンツ解析の研究に従事。電子情報通信学会会員, 博士(情報工学)。



松本 一則 (正会員)

1984年京都大学工学部情報工学科卒業。1986年同大学大学院修士課程修了。同年国際電信電話(株)入社, 研究所所属。現在, KDDI研究所知能メディアグループにて, マルチメディア

検索, コンテンツ配信の研究開発に従事。電子情報通信学会, 日本データベース学会各会員, 工学博士。



小野 智弘 (正会員)

1992年慶應義塾大学理工学部電気工学科卒業。1994年同大学大学院理工学研究科修士課程計算機科学専攻修了, 同年国際電信電話(株)入社。1999年9月~2000年9月スタンフォード大学電気工学科客員研究員。現在(株)

KDDI研究所知能メディアグループグループリーダー。この間, ソーシャルメディア解析, 利用者嗜好抽出, ソフトウェアエージェント等の研究に従事。電子情報通信学会会員, 博士(工学)。



東野 輝夫 (フェロー)

1979年大阪大学基礎工学部情報工学科卒業。1984年同大学大学院博士課程修了。工学博士。同年同大学助手。現在, 同大学大学院情報科学研究科教授。分散システム, モバイルコンピューティング, 通信プロトコル等の

研究に従事。情報処理学会フェロー, 電子情報通信学会, ACM各会員。IEEE Senior Member。