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†、服部元†、松本一則†、小野智弘†、東野輝夫††

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ロコミ情報を分析してマーケティング等に応用する評判解析技術に注目が集まっている。これらは手軽い情報発信が可能なため、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 上のロ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、ネット上のロ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。

Demographic Estimation of Twitter Users for Marketing Analysis

KAZUSHI IKEDA†, GEN HATTORI†, KAZUNORI MATSUMOTO†, CHIHIRO ONO†, TERUO HIGASHINO††

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.

† KDDI 研究所
KDDI R&D Laboratories Inc.
†† 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

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言われる[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 2)、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取っては「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。

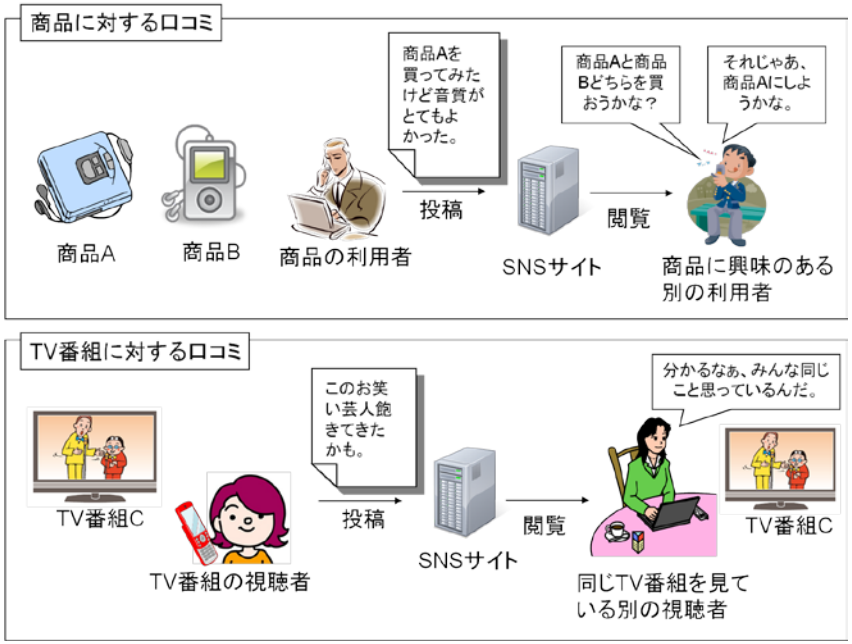

図 2 SNS における口コミ情報の共有

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

	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)	○

以降、2 章において提案手法の詳細について説明し、3 章では、提案手法の汎用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。4 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。

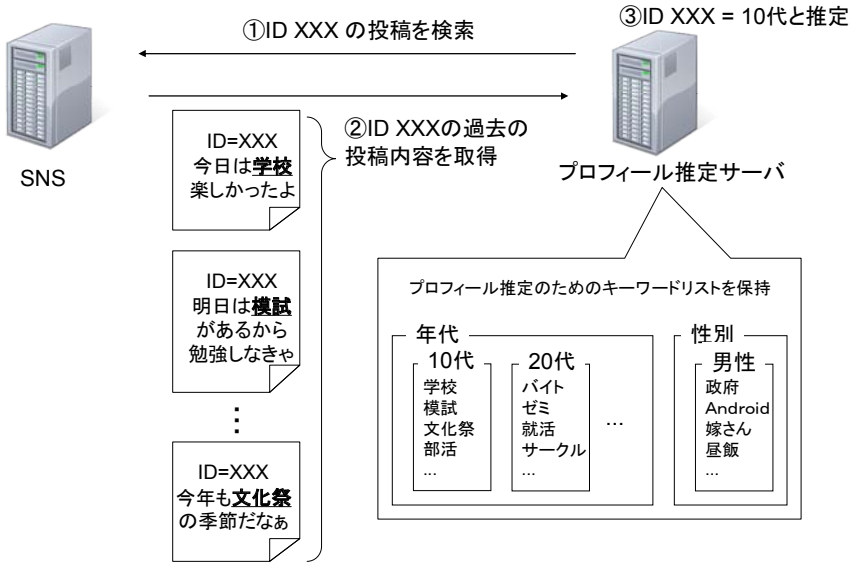

図 3 プロフィール推定技術の動作の仕組み

2. プロフィール推定技術

2.1 技術の概要

図 3 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、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) [5]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) [6]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。これらの詳細については、2.2 章、2.3 章でそれぞれ記述する。

2.2 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(例えば年代が 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。Twitter では、投稿者のプロフィールを自由文形式で記載することが可能であり、著者らの調査では年齢や性別の情報を記載している投稿者は少数であるが、それぞれ全体の 2〜3%程度存在した。これらのプロフィールが分かっている投稿者の過去の投稿内容を収集したものを SVM の学習用文書とする。

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)[5]を用いて算出する。表 2 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全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[7]の知見を元に、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。

$$
\begin{aligned}
&N_{11}(s) / N(s) > N_{12}(s) / N(\neg s) \ \text{のとき、} \\
&\quad E(s) = AIC\_IM(s) - AIC\_DM(s) \\
&N_{11}(s) / N(s) \leqq N_{12}(s) / N(\neg s) \ \text{のとき、} \\
&\quad E(s) = AIC\_DM(s) - AIC\_IM(s)
\end{aligned}
\qquad (1)
$$

ここで、AIC_IM(s), AIC_DM(s)はそれぞれ文献[5]の定義に従って、次の式で与えられる。

$$
\begin{aligned}
&AIC\_IM(s) = -2 \times MLL\_IM + 2 \times 2 \\
&MLL\_IM = N_p(s) \log N_p(s) + N(s) \log N(s) \\
&\quad + N_n(s) \log N_n(s) + N(\neg s) \log N(\neg s) - 2N \log N \\
&AIC\_DM(s) = -2 \times MLL\_DM + 2 \times 3 \\
&MLL\_DM = N_{11}(s) \log N_{11}(s) + N_{12}(s) \log N_{12}(s) \\
&\quad + N_{21}(s) \log N_{21}(s) + N_{22}(s) \log N_{22}(s) - N \log N
\end{aligned}
\qquad (2)
$$

表 2 E(s)値算出に用いる文字列 s の出現回数

	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(¬s)	N

表 3 文字列の出現回数と E(s)値の例

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

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 3 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)は負の値を取る。「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。

2.3 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 $S_1, S_2, S_3, \cdots, S_n$ 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 $M_1, M_2, M_3, \cdots, M_n$ 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 も合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全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、2.2 章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 4 SVM の入力となる特徴量の例

	S_1	S_2	S_3	…	S_n	Label (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

3. 性能評価実験

3.1 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 [8]、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[9]を用いた。実装には C 言語を用いた。

利用データ：Twitter のプロフィール欄に年齢、性別、居住地域のいずれかについて記載のある投稿者を対象に収集を行い、人手によってプロフィールごとに分類した。実験に利用した投稿者数を表 5 に示す。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)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 4)。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(推定に正解した投稿者数/全投稿者数)を評価、(2)視聴率測定などへの応用を想定し、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把握するため、与えられた投稿者のプロフィールの分布(均等、不均等それぞれ)に対して、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 $T_1,T_2,...T_n$(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 の真の投稿者数 $U_t(T_i)$、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。

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

$$E_{avg}(T) = \sum_{1}^{n}(|U_t(T_i) - U_e(T_i)|)/\sum_{1}^{n}U_t(T_i) \quad (3)$$

表 5 実験に利用したプロフィールごとの投稿者数(人)

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

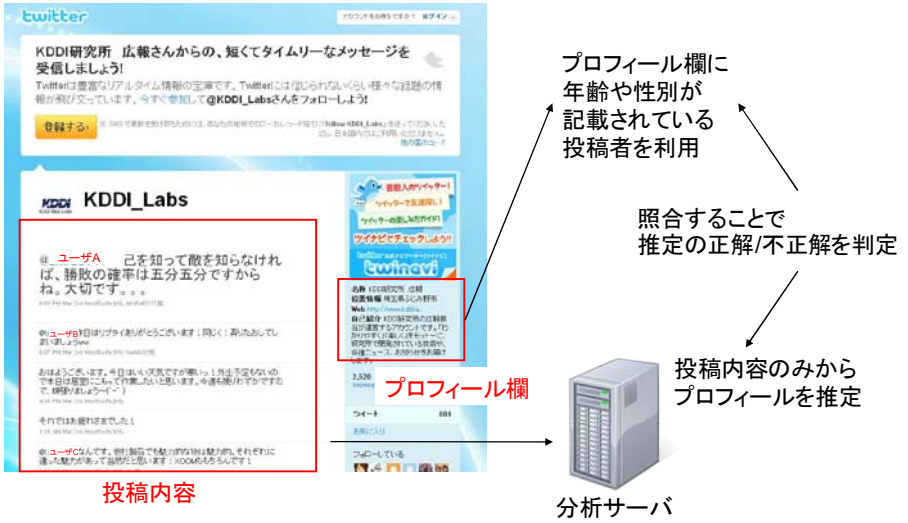

図 4 プロフィール推定精度の評価方法

(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。

表 6 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)

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	血圧	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	心斎橋
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

3.2 実験結果

表 6 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 7 に(1)のプロフィールごとの推定精度、表 8,9,10 に(2)の真の投稿者分布に対する推定された投稿者分布の比較、表 11 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 6 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、男女共に配偶者を指す言葉は特徴的と言える。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 7 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、女性共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いため、区別が困難なためと考えられる。

表 7 プロフィールごとの推定精度

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 8,9 から分布の均等、不均等共に、年代、性別、居住地域共に真の分布に近い推定結果が得られていることが分かる。表 7 の推定精度に比べて分布の誤差が小さいのは、表 7 では投稿者ごとに正解・不正解を判定しているが、例えば 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 10 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 11 はプロフィールごとの再現率、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%(推定結果は全て正解)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(例えば 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。

表 8 真の分布(均等)と推定された分布の比較

プロフィール	真の分布(均等)	推定された分布	誤差率(%)
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

表 9 真の分布(不均等)と推定された分布の比較

プロフィール	真の分布(不均等)	推定された分布	誤差率(%)
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 10 プロフィールの種類ごとの分布誤差平均 E_{avg}

	誤差平均(均等)	誤差平均(不均等)
年代	10.0	12.4
性別	5.4	8.8
居住地	4.6	14.0

表 11 プロフィールごとの再現率、適合率の関係

10 代		20 代		30 代		40 代	
再現率	適合率	再現率	適合率	再現率	適合率	再現率	適合率
25.1	100.0	4.2	95.0	2.2	50.0	4.0	66.7
40.0	98.9	13.1	93.7	6.4	63.0	10.9	71.0
49.1	97.4	21.8	91.6	11.6	64.2	19.3	69.6
54.9	95.4	28.7	89.6	16.0	55.4	28.8	69.5
...	...	...	...	...	...	...	...
83.8	83.8	68.0	77.3	50.7	55.1	69.4	57.9
男性		女性		北海道・東北		関東	
再現率	適合率	再現率	適合率	再現率	適合率	再現率	適合率
19.6	98.0	0.0	0.0	13.7	100.0	4.4	100.0
39.5	98.6	0.0	0.0	27.0	97.8	8.1	97.6
52.7	97.1	5.7	98.6	37.1	97.4	14.7	98.6
58.8	96.9	19.2	99.2	44.2	94.8	21.0	93.7
...	...	...	...	...	...	...	...
90.7	86.1	85.3	90.2	77.2	73.7	61.1	56.3
北信越		東海		近畿		中国・四国	
再現率	適合率	再現率	適合率	再現率	適合率	再現率	適合率
8.3	97.6	10.5	96.3	5.2	92.9	8.1	97.6
19.0	96.9	21.6	96.4	13.9	97.2	17.7	98.9
28.6	95.9	30.6	91.6	24.4	94.5	25.4	96.9
36.7	94.8	39.3	88.6	35.7	88.9	33.9	94.9
...	...	...	...	...	...	...	...
69.8	70.9	72.8	74.1	72.0	70.0	65.7	74.1

九州・沖縄	
再現率	適合率
18.1	97.8
29.2	98.0
39.1	94.2
47.4	91.1
...	...
76.2	77.3

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、居住地域の全てを推定しても 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と、1 台の PC のみを用いた場合でも、約 4 ヶ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、データベースに登録されていない新規ユーザが出現したとしても、遅延なくプロフィールを推定することが可能である。

4. 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ロ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目的で Twitter 上の意見を収集、分析して PC 画面等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 5)。

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 を元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、これらを合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 6)。

アプリケーション画面：アプリケーションの画面イメージを図 7 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。

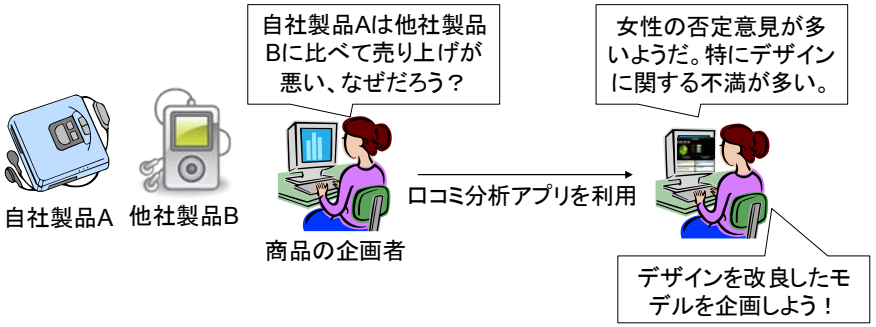

図 5 ロコミ分析アプリケーションの利用シーン

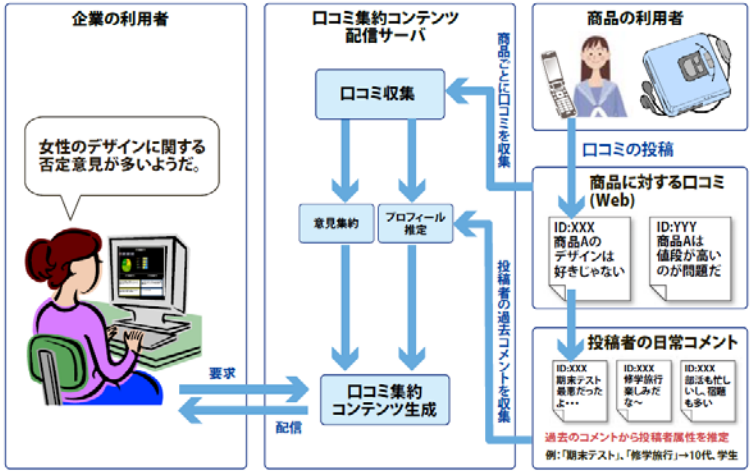

図 6 ロコミ分析アプリケーションの動作の仕組み

図 7 口コミ分析アプリケーションの画面イメージ

5. 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, http://blogs.itmedia.co.jp/saito/2011/02/mixi-twitter-fa.html
[4] K. Dave, S. Lawrence, D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” WWW, pp. 519–528, 2003.
[5] H. Akaike, “A New Look at the Statistical Model Identification,” IEEE Trans. on Automatic Control, Vol. 19, No. 6, pp. 716-723, 2003.
[6] C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, pp.273-297, 1995.
[7] K. Matsumoto and K. Hashimoto, “Schema Design for Causal Law Mining from Incomplete Database,” Proc. of Discovery Science: Second International Conference (DS'99), pp. 92-102, 1999.
[8] R. Fan, P. Chen and C. Lin, “Working Set Selection Using Second Order Information for Training SVM,” Journal of Machine Learning Research, vol. 6 pp. 1889-1918, 2005. (URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm/)
[9] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proc. of 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004) pp. 230–237, 2004.