ソーシャル・メディア・アナリティクス最前線
―センサーとしてのソーシャル・メディアとその分析技術―

ソーシャル・メディアを、世の中の動きを捉えることのできる一種の「センサー」として分析することで、世の中の関心事や会社への評判などを捉えることが可能となってきています。ソーシャル・メディアはテキストのほか、発信者の情報や時刻、ユーザーの行動など、多種多様な情報を含んでいます。これらの情報を活用するために、時間と話題（単語）の頻度と「誰が話をしているのか」「どの場所で話題が盛り上がっているのか」「どのように情報が流れていくのか」などを組み合わせた深い分析が重要となります。

本記事ではこれらソーシャル・メディアに特徴的な「テキスト分析」「情報の流れ」「ユーザー・プロファイル（人となり）」「ユーザー行動」の分析手法をそれぞれ紹介し、具体的にどのようなことが分かるのかを示します。

1. 新しいソーシャル・メディア分析の必要性

「ソーシャル・メディアの分析」というと、今までは「ソーシャル・メディア上のテキストの分析」を指すことが一般的でした。しかし、PROVISION76 号のコラム「日々是革新」[1] でもご紹介したように、ソーシャル・メディアは世の中の動きを捉えることのできる一種の「センサー」と見なすことができます。例えば、ソーシャル・メディア上のテキストを分析することで、世の中で話題になっている事柄や、製品やサービス、お店といった対象物に対する評判などを捉えることが可能となりました。

これらのテキストの分析では、ある表現（商品名や好評・不評など）の言及数を見るだけではなく、その時系列的な変化やその発言を発した人の偏りなどを見ることが重要です。このような技術は、テキスト・マイニングという分析で多くの成果が上げられています。もちろん、ソーシャル・メディアの分析において、テキスト・マイニングは大きな役割を果たしてきました。

しかし、ソーシャル・メディアには、テキストと発信時刻、ユーザー名の情報以外にも、さまざまな情報が含まれています。例えば、ソーシャル・メディア上でのプロファイル情報や過去の発言や行動などから、そのユーザーの人となりを知ることができ、それらを加味することでさらに深い分析を行うことができます。また、ソーシャル・メディアの重要な側面として、コミュニケーションのためのツールであるということが挙げられます。ソーシャル・メディアでは、通常の会話のような返答構造のほかに、Twitter のリツイート、Facebook のシェアといった、他のユーザーが書いた発言を再共有するという機能があり、ネット上での情報の拡散を促進しています。このような現象を捉えるためには、単なるソーシャル・メディアに投稿されたテキストのみを分析するのではなく、このようなソーシャル・メディアに特徴的な情報の流れやユーザー同士のコミュニケーションに関する情報を含めて分析することが必要となってくるでしょう。

本記事では、このようなソーシャル・メディアに特徴的な情報を用いた分析についての最新技術をご紹介します。

2. ソーシャル・メディアの 4 つの分析要素

ソーシャル・メディアではコンテンツの中心となるテキストのほかに、「情報の流れ」「ユーザー・プロファイル」、そして「ユーザーの行動情報」という 3 つの重要な分析の要素があります。

コンテンツとしてのテキストは、出現する名詞の頻度はもちろん、動詞や形容詞などの頻度や、それらの係り受けから、「何がどうした」というような文脈まで含めた情報を取得することができます。ソーシャル・メディアに対しては、これらの自然言語処理と呼ばれるテキスト分析を中心として、「情報の流れ」「ユー

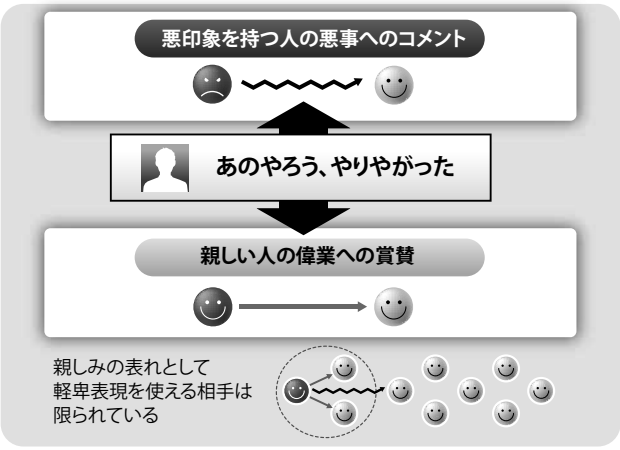

図 1. 軽卑表現を用いた炎上の判定

ザー・プロファイル」「ユーザーの行動」の要素を分析することが重要となります。次の章から、この４つの分析について、それぞれ詳しく触れていきます。

3. テキストを中心とした分析

<評判分析>

ブログなどの CGM（Consumer Generated Media：消費者が作り出すメディア）と呼ばれるコンテンツでは、商品やサービス、店舗などに対する率直な意見や感想が述べられています。これらの分析には、テキストの中の「好評」「不評」という評価表現を判断する「評判分析」という技術が用いられます [2]。

評判分析の一つの方法として、一つの文書に対して一つの評価（「好評」または「不評」）のラベルを付け、「好評が 20%、不評が 35%」などのように、文書の評価ラベルの割合を結果として出す方法があります。しかしこのような方法では、「軽くて持ち運びに良いが、充電池の持ちが悪い」といったような好評と不評の表現が混在する文章の内容まで正しく理解することができません。

そこでさらに深い分析を行うため、「軽くて持ち運ぶのに良い」「充電池の持ちが悪い」といった表現に分割し、各々に対して極性（好評・不評）を付与するようにします。これにより、製品やサービスに対する具体的な評価について知ることができます。これらの表現を抽出するためには表現ごとに辞書を用いる必要がありますが、同じ「長い」でも「充電時間が長い」では不評、「稼働時間が長い」は好評といったように、単なる用言（形容詞や動詞）の表現だけではなく、その対象物も理解しなければ正しく評価表現を分析することはできません。また用言が否定形かどうかで極性（好評か不評かどうか）が反転するため、辞書との単なる文字列の一致ではなく、深い言語処理が必要となります。

また、対象とする商品やサービスの分野（ドメインと呼びます）によって評価の表現が異なるという問題もあります。例えば「泣ける」という表現は、通常は不評の表現となりますが、映画の感想では好評と捉えることもできるでしょう。分析に当たっては、そのドメインの評価表現が多く含まれた文書を用いてドメイン依存の評価表現辞書を作成します。評価表現は「紫色がくっきりしていて、満足しています」というように好評→好評、あるいは不評→不評と同じ極性の表現が連続して現れるという現象があります。一方で「だが」「しかし」などの逆接の表現が現れると、「演技はうまくない。しかし、引き込まれた（不評→好評）」といった極性の反転が見られます。このような表現の「文脈一貫性」と「逆接による反転」を用いて、極性つきの評価を自動的に集めることで、ドメイン依存の評価表現辞書を作成することができます。

<軽卑表現を用いた炎上判定>

ソーシャル・メディアでは「炎上」と呼ばれる、ある特定の組織や人の行動や発言に対して非難が集中する現象があります。この「炎上」はこれらの現象がソーシャル・メディア上で拡散されたり、また既存メディアで報道されたりして、株価の下落や信用の失墜などソーシャル・メディアにとどまらず広く影響をもたらすことがあります。

「炎上」では、先に述べた評価表現だけではなく「〜しやがった」「このやろう」といった相手を非難する特徴的な表現を用いることが多く見受けられます。このような、相手に対して否定的な態度を示す表現を「軽卑表現」と呼びます。

ところが「あのやろう、やりやがった」のような表現は悪印象を持つ人への非難としてだけではなく、親しい人の偉業への賞賛に使われることもあります。しかし、**図 1** に示すようにこのような軽卑表現を親しみとして用いることのできる相手は、親しい間柄の関係に限定されます。親しい相手は人によって異なりますから、さまざまな人から同時に軽卑表現を向けられている人は非難の対象と見なすことができると考えられます。

また、ソーシャル・メディア上での「炎上」は、ある特定の組織や人に対して、非常に短い時間に起こるという特徴もあります。したがって、「ある特定の時間に集中してさまざまなユーザーから軽卑表現を用いられるかどうか」で、炎上の判定を行うことができます [3]。

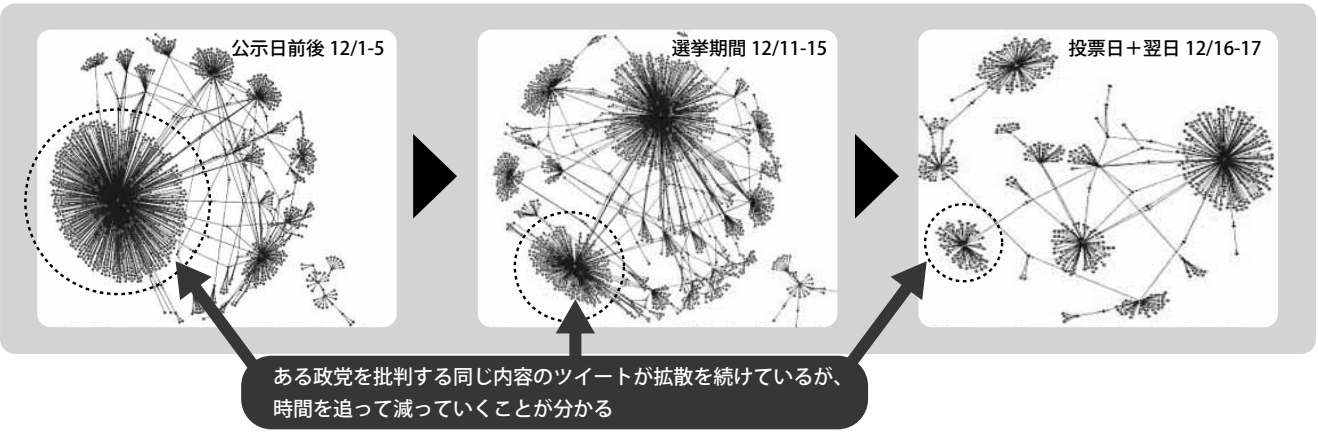

図 2. 公示日、選挙期間、投票日における「沖縄基地問題」に関連するツイートの拡散の様子

4. 情報の流れ分析

ソーシャル・メディアにおける炎上の現象の一つに、情報の拡散があります。情報の拡散とは、Facebook でのシェアや Twitter でのリツイートなどの「情報の再共有」によって、ソーシャル・メディア上での情報が広がっていく様子のことを言います。

図 2 に 2012 年 12 月に行われた衆議院選挙の公示日、選挙期間、投票日における「沖縄基地問題」に関連するツイートの拡散の様子を示します。さまざまな話題が拡散されていますが、その中で、基地問題を通してある政党を批判するツイートが継続的に拡散されていく様子が見て取れます。このように、選挙に関連するあるトピックに対して、どのような内容が拡散され、あるいは衰退していったのかを捉えることで、ソーシャル・メディアが選挙の結果にどのように影響を与えたかを考察する要素の一つとすることができるでしょう。

リツイートは本来、情報をそのまま伝達することで拡散します。しかし、本来の情報とは別の情報が付与されて拡散する例もあります。例えば、ある企業について悪い情報が発信されたとします。通常はその情報が共有されると、企業にとってダメージが大きくなります。しかし、その悪い情報に対する否定的な発言と共に再共有された場合、人々はその両方の情報を読むことになります。すると企業の悪い情報を覆すような、むしろ企業にとって良い方向に話が広まることもあり得ます。逆に、元は良い情報や中立の情報だったとしても、それが否定的な発言と共に広がった場合は企業にとってリスクになり得ます。したがって、情報の拡散する様子を単に捉えるだけではなく、それに対する意見も共に把握することが重要になります。

図 3 は、ある会社が新製品を発表した情報が拡散されていく様子を示したものです。当初は新製品の情報がそのまま拡散していきましたが、一人のユーザーがこの製品に対する否定的なコメントと共に再共有し、それも同時に拡散していく様子が見て取れます。これにより、その新製品に対する拡散の一部は、批判的な発言と共に拡散されたということを把握できます。

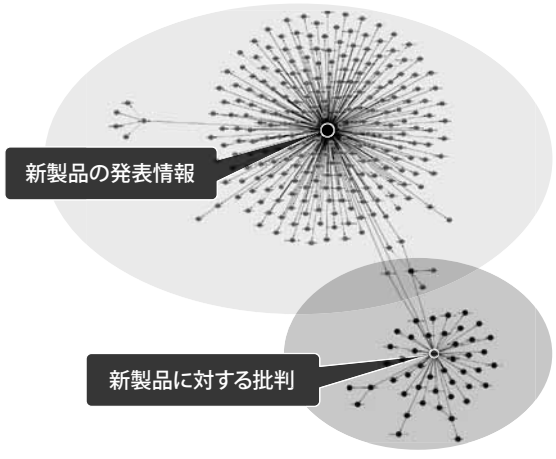

図 3. 元情報への意見情報を付与した拡散の視覚化

5. プロファイルの取得とそれを利用した分析

発言者のプロファイル情報はアンケートなどの分析では必須であり、年齢、性別、居住地域などの情報を用いて、どの年代でどの性別の人に特徴的な意見があるのかなどの分析が行われていました。

ソーシャル・メディアにおいても重要な要素の一つとして「誰が話をしているのか」という情報があります。ソーシャル・メディアのツールの中には「プロファイル情報」を提供しているものもあります。例えば、その中の「地域」の情報を利用すると、どの地域の人がある製品について多くつぶやいているか、といった情報を知ることができます。**図 4** に 2012 年の NHK 紅白歌合戦についての Twitter の分析を示します。活動の拠点から関西に人気の地域性があると期待される「関ジャニ∞（エイト）」は、実際は地域に偏りがなく全国的につぶやきがある一方で、大阪の路上ライブ出身で今でも関西に根強い人気がある「コブクロ」は、紅白関連のツイートでも関西地域の方のつぶやきが多いことが分かります。

このように、ソーシャル・メディアのテキストの分析と、それを書いたユーザーのプロファイルの分析結果との組み合わせから、地域性の偏りを代表とする多くの情報が得られると期待できます。しかし、すべてのユーザーがプロファイルを作成しているわけではないため、プロファイルを用いた分析を行うためには、そのユーザーの過去の発言などからプロファイルを推定してユーザーやその発言などの分析対象を拡大することが必要になります。

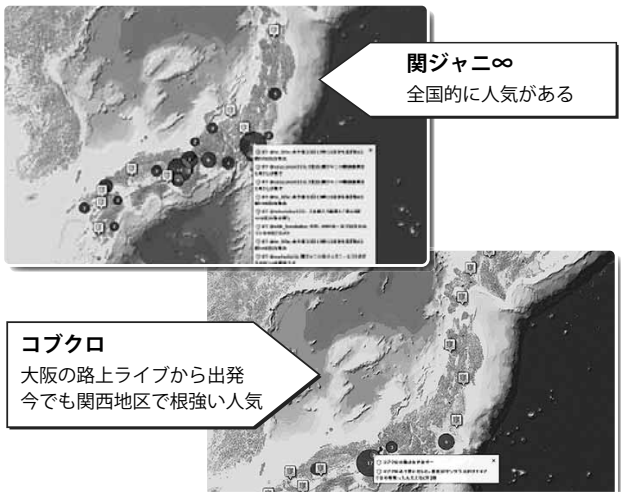

図 4. 紅白歌合戦におけるアーティストの地域性

<位置情報推定>

プロファイルの情報としての「場所」には、発言者の居住地やその発言を行った場所など、複数の観点があります。モバイル機器が発達した現代では、プロファイル情報に載せている居住地からだけではなく、旅行先といった居住地から離れている場所か

らの発言もあります。ソーシャル・メディアにおいては、発言の中で話題となっている場所、発言した場所、そしてその発言者の居住地という、大きく分けて 3 つの位置情報があると考えられます。

ある発言で話題となっている場所は、発言の中に含まれている地名などを抜き出すことで可能になりますが、発信場所、居住地は発言の中の地名を分析することでは把握できません。そのため、過去の発言などを蓄積して、居住地や発言場所を推定します。

居住地は、ユーザーの記述したプロファイル情報を用いればある程度推定することができます。しかし、例えば Twitter では、プロファイル中に実在する地名を居住地情報として載せているユーザーは 20 ～ 30%程度と言われています。そこで、あらかじめ居住地が分かっているユーザー情報を用いて、居住地がプロファイルに書かれていないほかのユーザーの居住地を推定します [4]。小規模な地震や通り魔などの犯罪、あるいは地元に密着したスポーツの勝敗などは、発言の地域に偏りがでることが想定されます。そのため、このような内容の発言をあらかじめ居住地が分かっているユーザーと同時に発言したユーザーは、その地域に居住していると推定することができます（**図5**）。

発言場所は、モバイル機器に搭載された GPS などによって発言に付与された「ジオタグ」や、その場所にいることを宣言する位置情報サービス・アプリケーション（例えばFoursquare [5]など）から生成された表現によって知ることができます。しかし、このような位置情報の付いたコンテンツは、全体の分析対象の数パーセントしかありません。そこで過去に取得した発言の中で、位置情報のある発言や時間的に近い発言を用いて、場所特有の表現を学習することにより発言場所を推定することができます [6]。

<過去の発言を用いたプロファイル属性分析>

プロファイルで書かれる内容は性別や年代、居住地域などが中心ですが、製品開発などで分析する際には、職業や子どもの有無、家や車といった所有物などのプロファイルが重要になります。

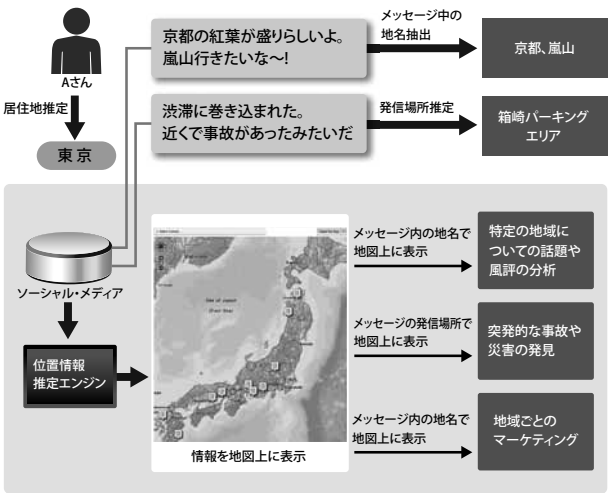

図 5. ソーシャル・メディアにおける位置情報の種類

例えば前章で紹介した評判分析などと組み合わせると、お子さんのいる女性に不評な機能は何なのか、スマートフォンを持っている人はこの製品を好むのかなどの、より詳細な分析が可能となります。

プロファイルを推定するためには、過去の発言で記述された通常のテキストから、「私の」「うちの」「僕の」といった一人称所有格表現に着目し、その表現が修飾する名詞句を参照します [7]。例えば、「私の夫」という表現を含むメッセージの発信者は配偶者を持つ女性であり、「うちの孫」という表現を含むメッセージの発信者は孫を持つ祖父母の世代であると推定することができます。ソーシャル・メディアで個人の属性をどの程度公開するかは、利用者本人に委ねられており、「子どもがいるか」「結婚しているか」は直接分からなくても、子どもがいる、結婚している可能性が高いユーザーを取得し、それを利用することでプロファイル付きの分析が可能になります。

6. 行動を対象とした分析

<閲覧率推定>

昨今のソーシャル・メディアには、自分が登録している友達の発言が、時系列に表示されるという特徴があります。リアルタイムに表示される形式になったことで、そのツールを利用しているときの発言しか読まれないという現象が起きています。そのため、フォローしている人の発言をすべて読んでいるとは想定できなくなりました。企業などが公式アカウントなどで発言するときは、より多くの方に読まれることを期待しますが、多くのユーザーにフォローされているアカウントであっても、そのフォロアーの多くがツールを使っていない時間帯に発言をした場合には、それらの発言はまったく読まれていない可能性もあることになります。

そこで、「ユーザーはいつソーシャル・メディアを利用しているのか」という、ソーシャル・メディア上でのユーザーの行動を知る必要が出てきます。ユーザーが発言した時刻を基本としてユーザーのログイン行動パターンを知ることができますが、ログインしたけれども発言していない場合などは、正しく行動を捉えることはできません。

そのため、発言の時刻だけではなく発言の種類を利用して、ユーザーのソーシャル・メディア上での行動パターンを推定します。ソーシャル・メディア上でのユーザーの発言は、自分の気持ちや状態を示した通常の投稿のほかに、他のユーザーに対する呼びかけや返答といったコミュニケーションが主体となる発言があります。多くのソーシャル・メディアは個人宛の発言は通常の発言とは別に表示させることができるため、自分宛ではない通常の発言よりも読まれる可能性が高いことが期待できます。また自分への呼びかけに対しては、その発言を読んでから時間をあけずに返答することが期待できますから、「呼びかけられてから返

答するまでの時間」は、ユーザーはそのシステムにログインしていなかったと推測できます。例えば**図 6** のようにユーザー B のユーザー A に対する呼びかけ「@ A ビール行きましょう」が発言されてから、ユーザー A が「@B 行きましょう」と返答するまで、25 分間の遅れがあったとします。発言しているときはログインしますが、一方でその発言からどれくらい前にログインしたのかは分かりません。しかし、A の B への返答がなされなかったこの 25 分の間は A がログインしていなかった、と仮定することができます。

ユーザーのソーシャル・メディアの利用に対する行動パターンは、一般的に、通勤時間や休み時間、寝る前にソーシャル・メディアを利用するなど、1 日のサイクルに依存していると考えられます。そのため、ユーザーの発言と先に挙げた反応への遅れの情報を積み上げることにより、ユーザーのソーシャル・メディアの 1 日の利用パターンを知ることが可能になります。**図 7** に、このようにして得られた 2 人のユーザーのソーシャル・メディア（Twitter）上での行動パターンを示します。縦軸がソーシャル・メディアの閲覧率、横軸が 1 日の時間（24 時間表示）を示しています。閲覧率とは、その時刻にそのユーザーに表示される発言が読まれる割合（1 であればすべて読む、0 であればまったく読まれない）を示しています。ユーザー①は朝 7 時台の閲覧率が比較的高く、あとは 24 時くらいまで 40% から 60% 程度を推移し、夜 24 時以降 5 時まではほぼ閲覧していないことが見て取れ、比較的、規則正しい生活を送っていることが分かります。一方、ユーザー②は、1 日中比較的高い閲覧率を保っており、夜も朝 4 時くらいまで閲覧が続くことが見て取れます。

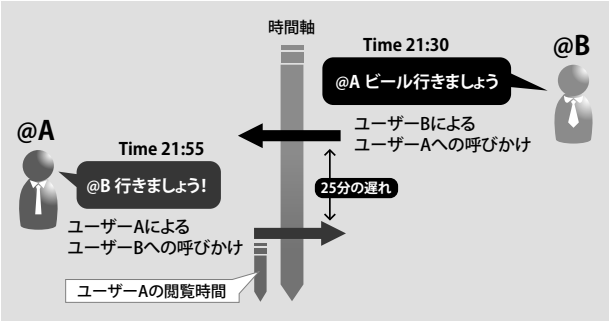

図 6. 返答によるユーザー行動の推測

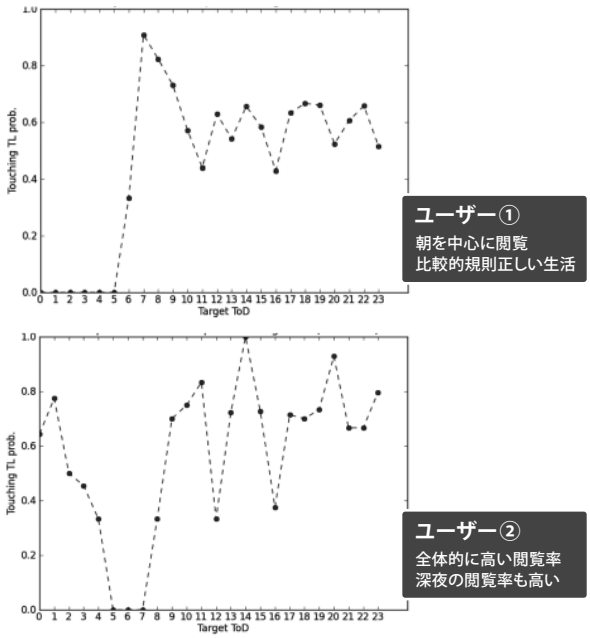

図 7. ユーザー 2 人の 1 日のソーシャル・メディア閲覧率

このようにして、ユーザーのソーシャル・メディアの閲覧率を取得できれば、自分の友達となっているユーザーの行動パターンを捉えることで、例えば発言するのに最適な時間（読まれるために一番効果的な発言の時刻）などを知ることができます。

<批判的ユーザーの推定>

ソーシャル・メディア上の返答などの行動はその人の交友関係を示すだけではなく、人や組織などへの支持も定量的に計ることができます。ネット上では支持を示すよりも、不快感や不支持を示す方が多い傾向にあります。例えば、あるアカウントに対して執拗に批判や非難を繰り返すなどの行為をするユーザーがいます。このようなユーザーがこの組織や人を定常的に批判するユーザーであることを事前に把握しておくことで、例えば定常的な批判と急激に起こった炎上を区別しやすくなります。また、このような批判的なユーザーの発言を定常的に追うことで、どのような内容が批判者を増やすのかという観点から現状の問題点や改善点などを把握することもできます。

批判をするユーザーは先に述べた軽卑表現を用いることが多い一方で、直接的ではない表現を使うこともあります。例えば「このご時勢に海外視察とはずいぶんと余裕があるのですね」などのいわゆる“嫌味”は、機械的に批判と捉えることは困難です。また、自身は批判的な発言をあまり行わずに、批判的な発言に対して同意をしたり拡散したりすることで否定的な立場を表すユーザーもいます。

そこで、批判的なユーザーを網羅的に取得するために、まず

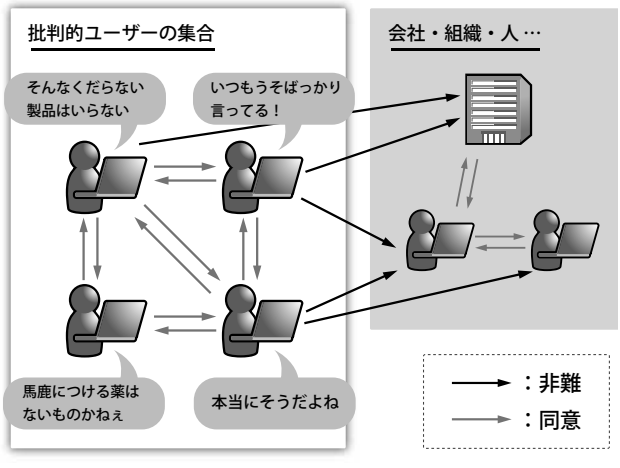

図 8. 発言内容と行動を元にした「批判的ユーザー」の発見

対象に直接的な批判をしているユーザーを言語的な分析を用いて取得します。さらに、そのユーザーの批判的意見に同意をしていたり、発言を拡散したりした発言を取得します。また、そのような直接的な行動を伴っていなくても、批判的ユーザーを多くフォローしているということは批判的ユーザーである可能性が高いとして、それらのユーザー間の関係も取得します。直接的な批判をするユーザーを基点とし、それらのユーザーに行動と関係で近しいユーザーを、最終的に批判的なユーザー集合として取得することができます（**図 8**）[8]。

7. ソーシャル・メディア分析活用に向けて

本記事では、ソーシャル・メディアの分析にどのような技術があるのかについて、東京基礎研究所での研究内容を中心に紹介してきました。これらの要素技術を用いることで、ソーシャル・メディアを「世の中のセンサー」として用いることができると筆者は信じています。

一方で、「意見を知る」ためのツールとして、ソーシャル・メディア分析をアンケート調査などと同じものとして扱うという方向性には危険があると感じています。例えば、2013 年よりネット上での選挙活動が認められ、ソーシャル・メディアを分析することにより投票行動の予測などができるのではないかと期待されています。現実を知るためにサンプリングをするという意味では、ソーシャル・メディアを分析することも、電話などによるアンケートの調査も同じです。しかし、アンケート調査では調査する対象の素性を指定してから、その中での傾向を可視化、分析しています。アンケートに回答する本人の自己申告にはなりますが、ある程度の信頼性は確保できていると言えます。一方で、ソーシャル・メディアは利用者の素性をすべて明らかにすることはできません。

また、**図 9** に示すように、インターネット・ユーザー、ソーシャル・メディア・ユーザーといったツールの利用層の偏りも影響してきます。さらには、匿名性や対外性（自分がどう見られたいのかを考えて発言する）の影響も否定できません。これらのユーザー層の偏りと発言への影響から、サンプル調査とソーシャル・メディア分析は同じ物として扱えないということが言えます。

しかし、ソーシャル・メディアでは誰かに強制された意見ではなく自発的に意見を述べているため、ある事柄（製品や会社、選挙における政党や候補者）に対して、より実直で素直な意見を語っている可能性があります。ここが、何かの目的を持って聞かれていることを意識してしまうアンケートとは異なるところです。

ソーシャル・メディア分析を行う際には、このような特性を考慮し、目的にあった分析を心掛けることで、ご紹介した分析技術をよりいっそう活用できると考えています。

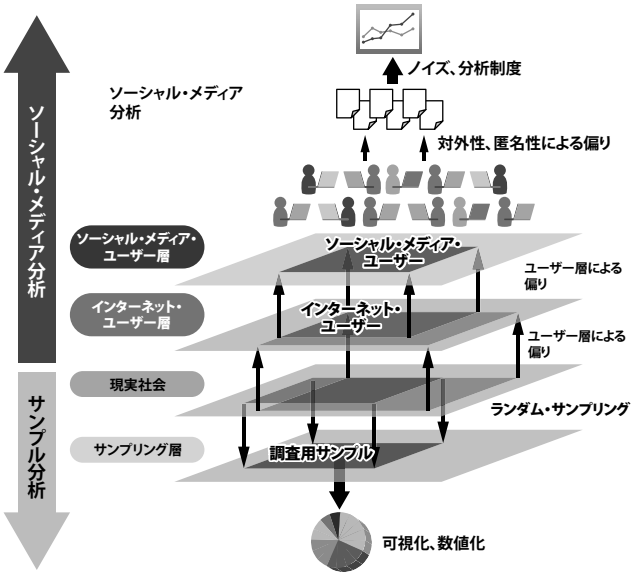

図 9. ソーシャル・メディア分析とサンプル分析との違い

［参考文献］
［1］東京基礎研究所「日々是革新」PROVISION76 号 ,「ソーシャル・メディア研究最前線 ソーシャル・アナリティクスで大量のデータから価値を」, pp78-80, 2013
［2］金山博 ,「テキストを用いた評判と嗜好の分析」, 情報処理 48（9）, pp1001-1007, 2007
［3］荻野 紫穂 , 那須川 哲哉 , 金山 博 , 榎 美紀 ,「軽卑表現の情報を活用した知識発見」, 言語処理学会第 18 回年次大会 , 2012
［4］山口祐人, 伊川洋平, 天笠俊之, 北川博之 ,「ソーシャルストリームからのイベント検出とユーザ位置推定の統合」, 第 5 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 2013）. 2013
［5］https://ja.foursquare.com/
［6］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）. 2012
［7］那須川哲哉 , 西山莉紗 , 金山博 , 吉田一星 , 大野正樹 ,「" 一人称所有格を用いたプロフィール推定」, 言語処理学会第 19 回年次大会 , 2013
［8］高瀬翔 , 村上明子 , 榎美紀 , 岡崎直観 , 乾健太郎 ,「ソーシャルメディア上の発言とユーザー間の関係を利用した批判的ユーザーの抽出」, 言語処理学会第 19 回年次大会 , 2013

IBM 東京基礎研究所（IBM Research-Tokyo）
アナリティクス&インテリジェンス
ナレッジ・インフラストラクチャ
リサーチ・スタッフ・メンバー

村上 明子
Akiko Murakami

［プロフィール］
1999 年、IBM 東京基礎研究所入社。以降、自然言語処理関連の研究に従事。テキスト・マイニング・ツールの研究開発において品詞管理や辞書作成などを担当した後、昨今では SNS（ソーシャル・ネットワーキング・サービス）などの人と人とのコミュニケーションを対象とする研究を行っている。近年では、IT を活用した災害からの復興や減災、リスク管理を実現する「レジリエント工学」の分野にも関わっている。著書として『チャンス発見の情報技術（東京電機大学出版）（共著）』、訳書として『情報検索の基礎（共立出版）（共訳）』などがある。