ビッグデータに対するテキストマイニング技術とその適用例

Text Mining Techniques for Analyzing Big Data

脇　森　浩　志

要　約　ソーシャルメディアや企業内の情報システムに蓄積されている大規模かつ更新頻度の高いテキストデータは“ビッグデータ”の代表例とされる．本稿では，企業活動において有益な知見をテキストデータから得るためのテキストマイニング技術の基礎，日本ユニシス独自の手法である“話題の変化を把握する技術”と“話題を可視化する技術”をその適用例とともに紹介する．ビッグデータの分析ではさまざまな情報を掛け合わせることでより大きな価値を生むとされるが，これらの手法は商品の売上や株価などの定量データとテキストデータを組み合わせて分析する際にも活用できる．

Abstract　Large volume text data that is frequently updated and stored in the enterprise information system and social media is a typical example of “Big Data”. This paper describes the basics of text mining, the techniques of deriving useful knowledge in business from text data. And this paper proposes a method of understanding the change of topics in text data and a method of visualizing topics in text data. In the analysis of big data, it is said that you create greater value by combining a variety of data. You can leverage our methods, when you analyze data that are combination of text data and numerical data such as stock prices and product sales.

1. は　じ　め　に

2011 年ごろから“ビッグデータ”を解析して企業活動に利用する取り組みが盛んになっている．ビッグデータとは Volume（規模の大きさ），Variety（多様な形式），Velocity（高い発生頻度）の特性から，従来の技術では管理・分析が難しいデータのことである．企業の情報システムには販売や接客などで発生するさまざまなデータが蓄積されている上，スマートデバイスやセンサーを備えた情報機器の普及が進んだことから大量のデータが手に入るようになり，その分析ニーズが高まっている．データマイニングや機械学習などの分析技術，高性能ハードウェアや並列分散処理などの基盤技術の進化により，分析のための土壌は整いつつある．

ビッグデータの対象となるデータには売上や株価といった定量情報のほか，テキスト，画像，動画，センサー，GPS，音声などが存在する．Twitter や Facebook，クチコミサイト，ブログといったソーシャルメディア上の書き込みはテキストデータとして蓄積され，ビッグデータの代表例として取り上げられる．本稿では，2 章にて企業活動におけるテキストデータの用途，3 章にてその分析技術について記述し，4 章では日本ユニシスの取り組みを紹介する．5 章は事例である．

ビッグデータの分析ではさまざまな情報を組み合わせることでより大きな価値を生むことができると考えられている．定量データの分析にテキストデータを組み合わせる技術についても併せて紹介する．

2. 企業活動に有益なテキストデータ

企業内外にはさまざまなテキストデータが存在する．本章では企業活動に有益なテキストデータとその用途を記述する．

2.1 テキストデータの種類と用途

企業活動に利用できるテキストデータには表1のようなものがある．

表1 テキストデータの種類

ソーシャルメディア	企業内
・Twitterのつぶやき ・SNSにおける発言 ・掲示板サイトの投稿 ・ブログ ・ECモールの商品レビュー	・コールセンターの応対記録 ・営業日報 ・商品アンケートの自由記述欄 ・Webコミュニティ上の投稿 ・顧客からの電子メール

ソーシャルメディアでは，自社商品や競合商品に対する意見や企業イメージ，CMへの反響，製品についての苦情がありのままに語られることから，表2のような場面で活用できる．

表2 テキストデータの用途

部門	用途
品質管理	製品やサービスの不具合を早期に発見
商品企画	商品開発につながる消費者ニーズの把握，競合商品との比較
広報	ニュースリリースやプロモーションの反響分析
接客	応対品質の改善
リスク管理	自社に対する風評の監視

TwitterやSNSでの発言は実世界をリアルタイムに映し出すため，社会環境の調査にも利用できる．例えばTwitter上で「風邪」を含むつぶやきを検索し，付随するGPS情報と合わせてその時間的な動きを見れば，風邪の時間的・空間的な広がりを観測できる可能性がある．議員選挙の当落予想や災害時の不足物資を推測する試みも行われている．

ソーシャルメディアが注目される一方で，従来から分析対象とされてきた企業内の情報も活用されている．コールセンターではソーシャルメディアと異なり，双方向のコミュニケーションが顧客との間で行われる．コールセンターの情報は，商品の不具合や顧客の要望を詳細に知ることができるため，品質管理や商品改善に欠かせない．また商品アンケートでは企業側が知りたい意見をピンポイントで収集することができる．意見の収集にソーシャルメディアを使うことで費用削減を図れるが，ソーシャルメディアでは語られにくい話題も存在する．必要に応じてアンケートを併用し，意見収集することが有効と考える．

ソーシャルメディア・企業内に関わらず，各々の媒体では発言者の属性や発言の目的が少しずつ異なる．分析時にはこうした点を考慮に入れて結果を解釈する必要がある．

2.2　定量データとの組み合わせによる効果

株価や売上，CS（顧客満足度）などは企業経営を大きく左右する重要な指標である．その好調・不調の原因を探るための参考情報として，ソーシャルメディアやコールセンターなどのテキストデータを役立てることができる．例えば，商品の売上や企業の株価が落ち込んだ際，その直前で商品や企業に対して悪評が広まっていれば，それに起因しているという仮説を立てられる．以後，ネガティブな話題の増加を早期に検知して原因を探ることができれば，信用失墜や賠償に伴う損害が大きくなる前に対策を立てることができる．

3.　テキストマイニング技術の基礎

テキストマイニングとは大量のテキストデータの中から価値のある知見を得る技術のことである．本章ではテキストマイニング技術の基礎について記述する．

テキストデータは定量データとは異なり文字列であるため，そのままでは扱いづらい“非構造化データ”である．構文を解析して単語や係り受けを抽出するか，内容に応じた分類付けをしないと集計や分析に利用できない．一般にテキストからの単語抽出は“形態素解析”，係り受け抽出は“係り受け解析”という技術によって行われる．

3.1　形態素解析とセンチメント分析

テキストデータに含まれる文を単語に分割し，その品詞を特定することを形態素解析と呼ぶ．日本文は，英文のように文中の単語が空白で区切られていないため，他国語より難解である．そのため，より精度を高めようと盛んに研究が行われてきた．主流となっているのは，大量の文を人手で単語に分割した正答データ（コーパス）を用いて，各単語の出現のしやすさ，品詞同士の隣り合いやすさをスコア化して“辞書”として保持し，解析時に利用する手法である．この辞書には単語のバリエーションがスコアと共に保持される．

多くのテキストマイニングツールでは，文に含まれる単語のランキングを表示して全体を概観する．この時，図1のように各単語がポジティブまたはネガティブな意味を持つのかを示し，内容をつかみやすくするツールも存在する．

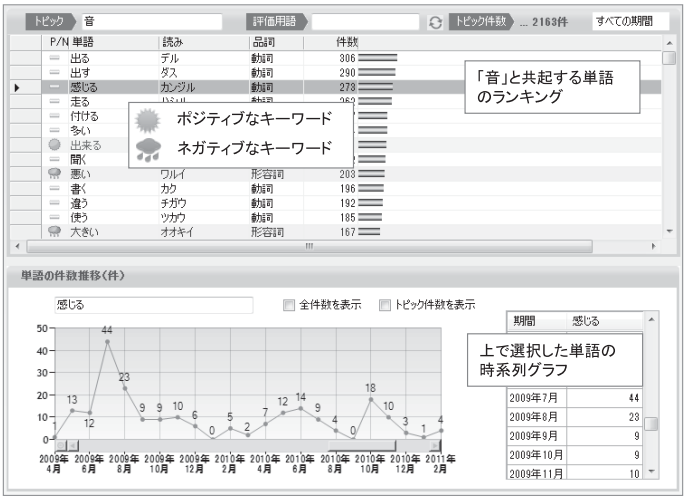

図1　TopicExplorer における単語ランキング

各単語のポジティブ・ネガティブの判定は“センチメント分析”と呼ばれ，文中の単語の共起関係を利用する方法，類義語や同義語の語彙集（シソーラス）を利用する方法などで行われる[1]．本稿では前者について解説する．前者ではポジティブあるいはネガティブの典型的な語をあらかじめ用意しておいて，どちらの語と共起しやすいかに基づき単語を選別する．共起の尺度としては以下の PMI（Pointwise Mutual Information）[2][3]がよく用いられる．式の p(A) は単語 A が出現する確率，p(B) は単語 B が出現する確率，p(A, B) は単語 A と B が共に出現する確率である．

$$PMI(A, B) = \log \frac{p(A, B)}{p(A)p(B)}$$

ある単語のポジティブ・ネガティブの判定はそれぞれの典型的な語との PMI を求めて，その値が大きい方とするなどの方法で行う．この時，両者の PMI に大きな差がない単語はどちらでもないと判定する．

3. 2　係り受け解析

主語–述語，修飾語–被修飾語など文節間の依存関係のことを係り受けと呼ぶ．文から係り受けを抽出する手順は以下のとおりである．

Step1.　形態素解析によって文を単語に分割する

Step2.　隣接する単語の前後で文節の切れ目を判定する

Step3.　文節を二つずつ取り上げ，係り受け関係の有無を判定する

文節の判定，文節間の係り受けの判定には SVM（Support Vector Machine）を用いて行う方法がある[2][4]．SVM には手作業で作成した文節や係り受けの正答データを学習させており，そこから導き出された規則を用いて，結果を推定する．

テキストマイニングツールでは単語と同様に係り受けをランキングとして表示し，全体を概観する．また，図 1 の単語ランキングにおいて単語だけでは内容の解釈が難しい場合に，その単語を含む係り受けの一覧を図 2 のように表示することで利用者の理解を助ける．

係り受け	件数
危険 − 感じる	14
私 − 感じる	13
大きい − 感じる	7
違和感 − 感じる	7
音 − 感じる	7
疑問 − 感じる	6
うるさい − 感じる	6

図1の単語「感じる」に着目

図 2　単語「感じる」を含む係り受けの一覧

3. 3　文書分類

テキストデータに対して自動で内容に応じた分類付けをすることを“文書分類”と呼ぶ．企業のコールセンターではデータ登録時に人手で分類が付与されることがある．しかし，ほとんどのソーシャルメディアでは書き込み時に分類は付与されない．集計・分析時に人手で対応することもできるが，対象が数万件に及ぶ場合，困難である．そこで文書分類の技術が必要とさ

れる．文書分類には，既に分類付けしてあるデータを用いて自動で付与ルールを作成する方法と，人が知識や直感によって付与ルールを作成する方法がある．文書分類の身近な応用例として電子メールソフトの自動仕分け機能がある．電子メールソフトでは，件名や本文に含まれる文字でメールを特定のメールフォルダーに仕分けするルールを登録することができる．また，スパムフィルタは不要なメールを自動で判別してくれる．このスパムフィルタは一般にナイーブベイズ分類器[2]を用いており，人がスパムと判定した過去メールを学習している．

図3のように分類ごとの件数を集計することで話題の構成比を把握できる．単語のランキングと比べ，業務に沿った視点で情報を整理することができるため，企業の管理層に対するレポートに向いている．しかし分類体系の整備や付与ルールの作成といった準備に多くの時間が掛かる．またソーシャルメディアのように新たな話題が次々と挙がってくる媒体では，分類を継続的に見直す必要があるため，維持が難しく不向きである．

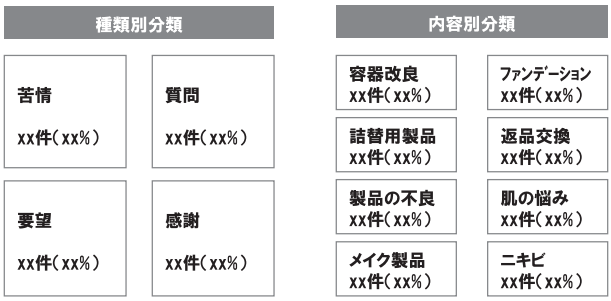

図3　文書分類したデータの集計例

4.　日本ユニシスのテキストマイニング技術

日本ユニシスでは1997年以降，テキストマイニングの製品開発およびその適用を行ってきた．本章ではその中で培ってきた，テキストデータで語られている話題の変化を把握する技術，話題の内容を可視化する技術について記述する．ともにビッグデータの特性であるVolume（規模の大きさ），Velocity（高い更新頻度）に対応した独自アプローチを採っている．

4.1　話題の変化を把握する技術

テキストデータで語られている話題の経時的な変化を把握する手法について紹介する．ここで“話題の変化”とはさまざまな話題の出現件数の増減を指している．

クチコミサイトにおける話題の増減から消費者トレンドの変化を早期に捉えることができれば，他社に先駆けてニーズにあった商品を提供し，大きな収益をあげることができる．また話題の増減は新商品の発売やTV番組・CMによる情報発信，製品不具合の発生などに起因することが多いため，2.2節で述べた定量データと組み合わせた分析の際に有効な情報となる．

4.1.1　話題の変化を把握するには

話題の変化を捉える一般的な方法は，単語や係り受けの出現件数を時間ごとに集計してその推移を確認することである．例えば，特定の時期に「危ない」「故障」などの単語が突発的に増えていれば，製品欠陥に関わる話題が急騰していると推測できる．

この方法では，着目すべき単語や係り受けが決まっていれば，人手で傾向を捉えることができる．しかし分析テーマに合わせて単語や係り受けを適切に選ぶにはノウハウが必要な上，大きな手間が掛かる．また，選んだ単語や係り受けに漏れがある場合，重大な事象を見落としかねない．そこですべての単語や係り受けを対象にして自動的に変化を捉え，通知してくれる仕組みが必要となる．

図4は，ある自動車のクチコミデータにおける単語「危ない」の出現件数を表した時系列グラフである．これに対して複数の直線からなる傾向線を当てはめる．

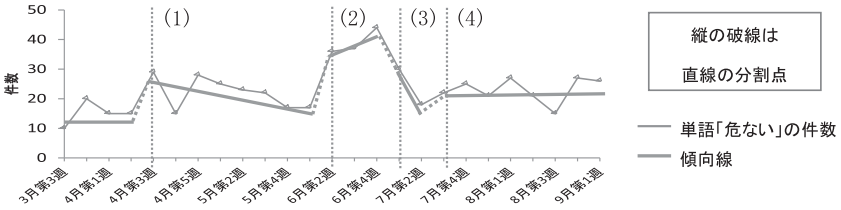

図4　単語「危ない」に対する傾向線の当てはめ例

図中(1)～(4)のような直線同士のつなぎ目を分割点と呼び，各直線の傾きや分割点で隣りあう直線の関係を分析することで，以下のような特徴を把握する．

［急増］　　特定の期からその次期で急に増加する
［増加傾向］　　特定の期から数期に渡って継続的に増加する
［減少傾向］　　特定の期から数期に渡って継続的に減少する
［急減］　　特定の期からその次期で急に減少する

この処理をすべての単語に適用する．見つけた増加傾向または減少傾向は，直線の傾きから1期あたりの伸び数や伸び率を求める．同様に，急増点や急減点についても前期からの伸び数や伸び率を求める．伸び数や伸び率を手がかりとして，変化の大きい単語に着目すれば，話題の変化を類推できる．

なお，傾向線を当てはめる際には，変化への“感度”を調整するパラメタw（正の整数）をあたえる．パラメタwは図5のように大きな値にするほど変化の捉え方がおおまかになる．

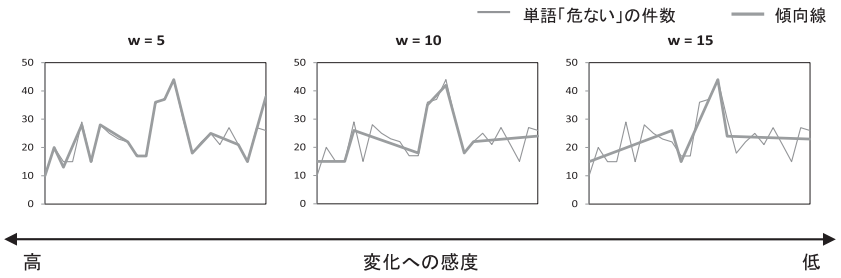

図5　単語「危ない」に対する感度パラメタwと傾向線の関係

一概に話題の変化を捉えるといっても，利用者は大まかな傾向だけを捉えたい場合もあれば，細かい増減まで知りたい場合もある．例えば，図4の(4)の右側では，全体を平坦とする

見方もあれば，8 月第 3 週の付近で減少しているという見方もある．こういった小さな変化への“感度”をパラメタ w によって調整できる．

4. 1. 2　話題の変化を把握する具体的な手法

4. 1. 1 項で示した機能を実現する具体的な手法について詳細を記述する．

1）　傾向線の当てはめ方法

図 6 を使って，ある単語の出現件数を表したグラフに傾向線を当てはめる方法を示す．この方法では図中の Step1，Step2 の手順で傾向線を作り上げる．

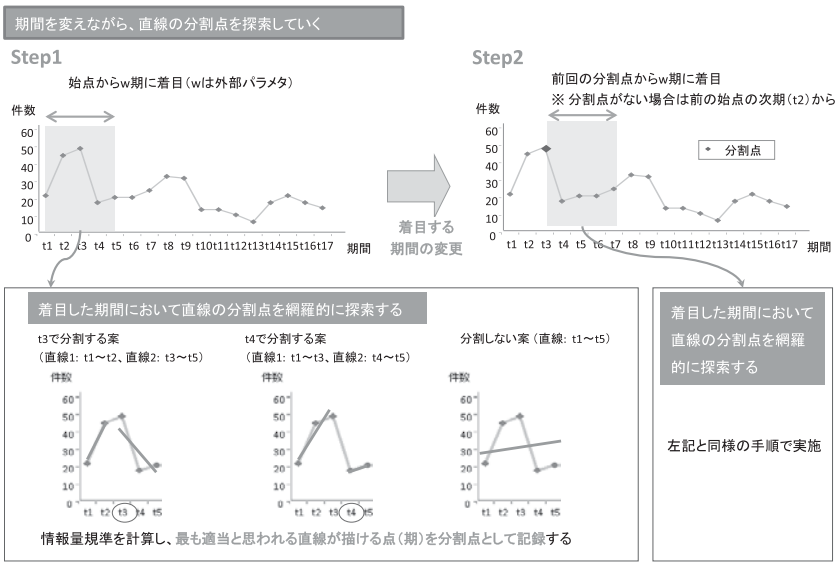

図 6　傾向線の当てはめ方法

Step1）

まず，時間軸（横軸）の左端を始点として w 期の点を取り上げ，この中で直線の分割点を探すことを考える．そのため，取り上げた点を前から順に着目し，その前後で回帰直線を引いてみる．すべてのパターンが網羅できたらそれぞれ“情報量規準”を計算し，その値が最も小さい点を分割点とする．ただし，w 期を 1 本の回帰直線で引いた場合より情報量規準が大きい場合には分割点を設けない．情報量規準とは統計モデルの当てはまりの良さを評価する指標である．この値を用いることで，出現件数と傾向線との誤差を抑えつつ，直線の分割が多くなりすぎないように調整することができる．情報量規準には赤池情報量規準（AIC）などを採用できる[5]．

Step2）

次に，直前に分割点とした点から w 期の点を取り上げ，Step1 と同様に分割点を探す．もし直前に分割点を設けなかった場合には前回の始点の次期を始点とする．以降，始点がデータの末尾になるまで繰り返す．分割点が決まったら，その前後で回帰直線を分けて引くことにより傾向線を作り上げる．

時間軸の点がｔ個あるとすると，直線や曲線の分割パターンは2^t通りとなる．期間が長くｔの大きいデータでは，すべてを網羅して比較するには膨大な計算が必要となる．それに対して本手法は，組み合わせを効率よく絞ることによって，計算時間をｔに比例するように軽減し，指数的な増加を防いでいる．

2）　新たなデータの追加に伴うアウトプットの更新

新たなデータが頻繁に追加される場合には以前のアウトプットを効果的に利用し，計算量を抑制する必要がある．1）で述べた通り，傾向線の分割点は，まずは時間軸の左端からｗ期の点を取り上げ，次第に対象を右に移して探索する．新しいデータが追加されると右端に点が加わるが，分割点を探索した結果は，直近ｗ期より前の期間では変わらないため，それを再利用できる．データ追加時には，直近ｗ期より前の分割点の中で最も右側にあるものを始点とし，ｗ期の点を取り上げるところから始める．後は同じように対象を右に移しながら分割点を探していく．

3）　複数の感度パラメタｗの指定による精度の向上

4. 1. 1 項で述べた感度を調整するパラメタｗは一度に複数あたえることができる．感度パラメタが複数の場合，まず，各値を使って，1）の方法により傾向線を当てはめる．次に，それぞれ情報量規準を計算，最も小さいものを最適な傾向線として，急増・急減などの特徴を判別する．感度パラメタｗを複数あたえることで，多くの傾向線の当てはめ方を候補に入れることができる．そのため，より納得性のある結果を得ることができる．一方で，計算量はあたえるｗの数に比例するため，所要時間は増加する．納得性と所要時間のバランス調整が必要となる．

4. 2　話題を可視化する技術

4. 1 節の手法を適用すれば，特定の単語の増加を検知することができる．次に利用者はその単語が文脈の中でどのように使われているかを知り，変化している事象を把握する．これは本節で紹介する“単語マップ”によって実現することができる．単語マップは図 1 に示した単語ランキングにおいて，特定の単語について語られている内容を把握する場合にも利用できる．

4. 2. 1　単語マップとは

図 4 に例示した「危ない」という単語が急増した 6 月第 2 週に着目して，単語マップを描くと図 7 のようになる．

図 7 の左側のように，このマップ上では着目した単語「危ない」の周辺にその単語と関連するキーワードを表示する．各キーワードからは，図 7 の右側の(1)～(3)のように，さらにキーワードを展開させることができる．例えば，(1)では「危ない」と「静か」の両方と関連するキーワードを展開している．この機能により，話題を掘り下げて把握することが可能となる．

図中(1)(2)からは「エンジン音がしないため歩行者は危ない」という話題が，(3)からは「雨の日は危ないので安全運転する」という話題が類推できる．最終的に(1)では，「危ない」と「静か」がともに出現する原文を参照することで類推の正しさを裏付ける．

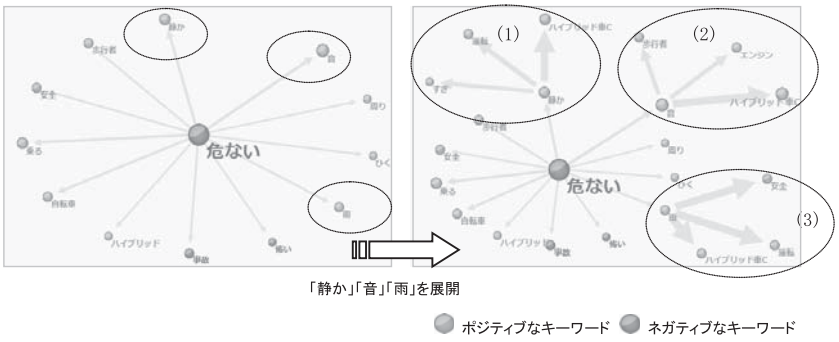

図 7　単語マップによる話題の可視化

4. 2. 2　単語マップの実現方法

4. 2. 1 項で示した手法について，その実現方法を記述する．

1)　キーワードの選別

助詞や助動詞などは“機能語”と呼ばれる[6]．機能語は文法的機能を目的とする語のことで，単独では意味を持たない．そのためマップ上ではすべて除外する．一方，名詞や動詞の中にも「もの」「こと」「する」など，どの文にも現れ，文意の把握に役立たない一般語が存在する．これらの単語もマップ上に表示しても不要な情報となるため除外する．

2)　キーワードの効率的な探索

ある単語の周辺に表示するキーワードは，その単語と共に出現する頻度の高いキーワードとする．キーワードを何回も展開している場合は，展開元を最上位まで遡って，すべての単語と共起する頻度の高いキーワードを表示する．例えば図 7(2)の「エンジン」を展開する場合には「エンジン」だけではなく「音」「危ない」が対象となる．

展開するキーワードには出現頻度による足切りを設けることで，ボリュームのあるテキストデータに対して効率的な探索を実現している．これについて詳述する．

キーワード A が出現するテキストの集合を {A} と表記する．また，キーワード A とキーワード B が共に出現するテキストの集合を {A} ∩ {B} と表記する．この時，キーワード K1 から展開したキーワード K2 を展開する場合には {K1} ∩ {K2} ∩ {X} の件数 ≧（足切りの出現頻度）を満たすキーワード X を探せばよい．ここで {K1} ∩ {K2} ∩ {X} は {K1} ∩ {X} の部分集合であるため，{K1} ∩ {X} の件数 ≧（足切りの出現頻度）も必ず成立する．この式はキーワード K1 を展開した時の条件式なので，キーワード X は K1 の展開時にも必ず対象となっている．

以上を一般化して考えると，あるキーワードを展開する場合は，その展開元のキーワードで対象となったキーワードのみを対象として調べればよいことが分かる．なお，図 7 では 1 キーワードから展開するキーワードの数を最大 10 件に絞っているため，(1)～(3)のすべてのキーワードが，展開元である「危ない」の周辺に表示されているわけではない．表示を絞っているのは，利用者が一目で内容を把握できるように重要な情報に限定するためである．

5. 適用事例

4章で記述した手法は日本ユニシスのテキスト分析ソリューション TopicExplorer および TopicStation に採用されている．本章では両ソリューションでの活用例について紹介する．

5.1 製品やサービスの改善点を見つける

製品やサービスに対するクチコミ中で急騰する話題は，消費者がその製品やサービスに求めているポイントに関係することが多い．図8は「ネットバンキング」に対して語られたツイート（Twitter 上のつぶやき）の件数と本手法で検出した急増キーワードである．

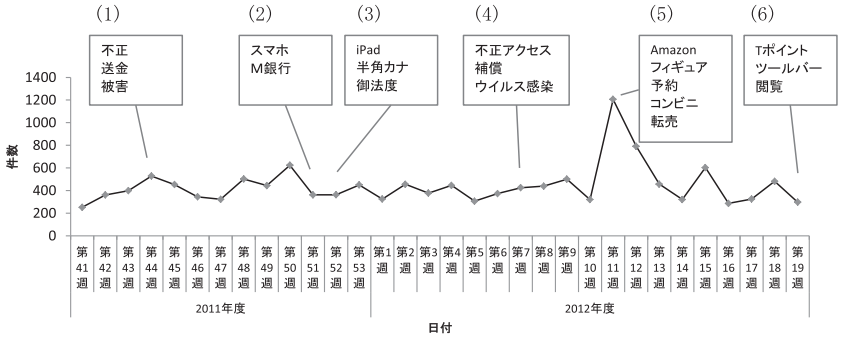

図8　ネットバンキングに対するツイートと増加キーワード

図中(1)～(6)のキーワードが現れるツイートの例文は表3のとおりである．(1)(4)(6)の期間ではセキュリティ面での不安が語られている．また(2)(3)ではスマートフォンからも利用したいという要望が，(4)では不正アクセスに対する補償をして欲しいという要望が挙がっている．これによりネットバンキングサービスを提供する銀行は，インターネット犯罪に対してのセキュリティ強化や補償を充実させつつ，スマートフォンでも利用できるようにしてほしいという消費者の期待に早く気付くことができ，自社サービスの改善に活用することができる．

表3　急増キーワードが出現する原文の例

期間	代表的な話題
(1)	ネットバンキングで預金が不正送金される被害が相次いでいることを受けて、…
(2)	M銀行、法人向けネットバンキングのスマートフォンサービスを4／2開始
(3)	ネットバンキングでiPadから半角カナを打つ必要があるが、どうすれば …
(4)	H銀行の法人口座にご注意！…ネットバンキングに不正アクセス…補償がありません
(5)	【Amazon】フィギュアの予約で… ネットバンキングが利用できなくなる？
(6)	Tポイントツールバーって、WEB閲覧履歴抜かれるって事は「どこの銀行でネットバンキングしてるか」とか … とかまで抜かれるって事ですよね？

5.2 消費者のトレンドを理解して商品開発に活かす

クチコミサイトに投稿された評価コメントのキーワードの変化を探ることで，消費者のトレンドを捉えることができる．表4はクチコミサイトのフレグランス（香水）ジャンルにおいて，2011年時点で増加傾向または減少傾向にあるキーワードの抜粋である．

表 4　評価コメント中の増加・減少キーワード

増加傾向	減少傾向
上品	セクシー
清潔	スパイシー
石鹸	キュート
持続	

「セクシー」「スパイシー」などといった個性の強い香りより，「上品」「清潔」といった清楚な香りを持つ香水に注目が集まっていることが分かる．また，香りの「持続」に対しての関心が高まっていることが類推できる．このクチコミサイトでは評価コメントに加えて，投稿者の年齢や商品に対する評点（1 ～ 10 点）が掲載される．TopicExplorer ではこれらの属性と出現件数の推移を図 9 のようなグラフで確認できる．

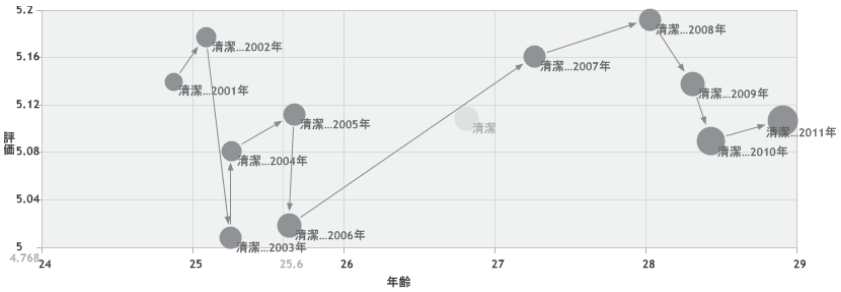

図 9　評価コメントにおける「清潔」の経時的な変遷

グラフでは「清潔」について着目している．横軸に年齢，縦軸に評点をとり，各年の平均値を円としてプロットし，矢印で結んでいる．プロットした円のサイズは出現件数に比例して大きくなる．このマップから「清潔」の出現件数が年ごとに多くなっていることを確認できる．また，投稿者の年齢が次第に右側に移ってきており，「清潔」な香りに興味を持つ消費者層が高い年齢にも拡大していることが類推できる．こういった洞察はトレンドを反映した商品開発に有効と考える．

5. 3　株価や売上不振の原因をクチコミから分析する

企業の株価や特定の商品の売上が浮き沈みするとき，その予兆として，企業や商品に対するポジティブな意見，ネガティブな意見がソーシャルメディア上で増え始めることがある．これを利用して，その好不調の原因を分析する．

図 10 は N 社の株価と日経平均株価の推移である．企業の株価は一般に，その企業にとってプラスまたはマイナスな要因がなければ，日経平均株価に連動するといわれているが，図中の (1) (2) の期間は異なる動きをしている．

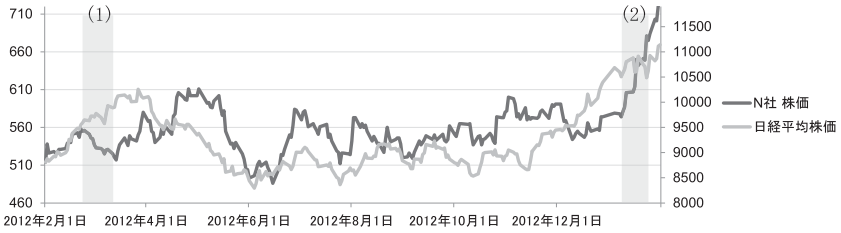

図 10　日経平均・企業の株価の比較

本手法によって各期間の前後で N 社に関するツイートで増え始めたキーワードを調べると，(1)では「AIJ」，(2)では「提携」というキーワードが増えていた．これらについて話題の可視化を行った結果を図 11 に示す．

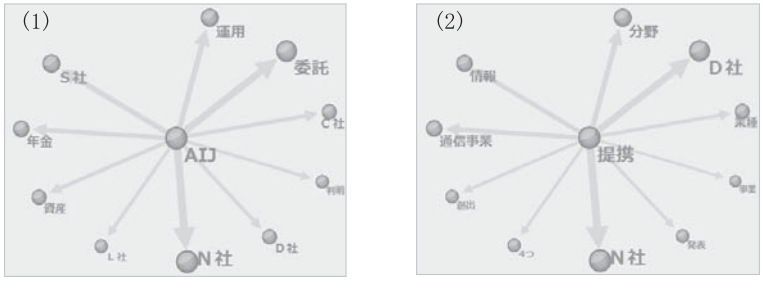

図 11　Twitter 上で急騰したキーワード

(1)では「年金運用委託先 AIJ の問題」についての話題が Twitter 上で急騰していることがわかり，企業の株価に影響をおよぼしたという仮説を立てられる．一方で，(2)では「D 社と異業種提携で新事業を創出」していくという明るいニュースがあり，株価の上昇に貢献したという仮説が立てられる．

6.　テキストマイニング技術の今後の展望

商品やプロモーションを新たに打ち出す際にはターゲットとする客層を定める．2. 1 節で挙げた各媒体において発言者のプロフィールが分かれば，投稿内容と合わせて分析することで，商品やプロモーションが実際にターゲットとしていた客層に受け入れられたのか，どのような点で不満を抱かせたかを把握することができる．また，想定と異なる客層に人気がある場合には，企業が商品やプロモーションの思いもよらなかった長所を知ることができる．

SNS では年齢や性別，趣味など，発言者側で公開しているプロフィール情報を取得することができる．一方，Twitter やクチコミサイトなどではこれらの情報はほとんど得られない．しかし発言に含まれる言葉や GPS の情報によってある程度の推測は可能である．例えば「娘にプレゼントを買った」という投稿があれば，既婚者かつ子持ちであることを推測できる．併せてその他の投稿に付随する GPS 情報で，深夜の所在地が新宿近辺に偏っていることがわかれば，その付近に在住と推測できる．性別や年齢には発言における言葉遣いを判断材料にできる．今後はこうしたプロフィール推定の技術が企業内でのデータ活用で求められると考える．

7. お わ り に

総務省と経済産業省は国や地方自治体などの行政機関が持つデータを情報システムから利用しやすい形で公開する準備を進めている[7]．公開対象には気象情報や交通情報など，企業内の情報と組み合わせることで新たな価値を生みだすことが期待できるデータも含まれている．行政機関が持つ豊富な統計データが加わることにより，分析の幅はさらに広がる．その一方で，分析の難易度も上がることになり，分析者に求められるスキルは一層高くなる．

4.2節で紹介した“話題の変化を把握する技術”はテキストデータだけでなく，定量データの変化も捉えることができる．例えば，生産ラインにおける不良品率を監視し，その異変を通知するといったことにも応用できる．同様に5章で取り上げた企業の株価や商品の売上も監視できる．将来的には，これらの異変を検知してその原因をソーシャルメディア上のクチコミから探るといったことが半自動的にできる可能性がある．これに関してはさらなる研究を進め，実用化したいと考えている．今後もさまざまなデータの相互の関係を自動で検出する技術や分析ツールを提供することで，分析者の負荷を軽減していきたい．

参考文献
［1］ 大塚裕子，乾孝司，奥村学，意見分析エンジン―計算言語学と社会学の接点―，コロナ社，2007年10月
［2］ 奥村学，高村大也，言語処理のための機械学習入門，コロナ社，2010年8月
［3］ Peter D. Turney, “Thumbs up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, In Proceedings of ACL, pp.417-424, 2002
［4］ 工藤拓，松本裕治，チャンキングの段階適用による日本語係り受け解析，情報処理学会研究報告，Vol.43 No.6，pp.1834-1842，2002年
［5］ 下平英寿，伊藤秀一，久保川達也，竹内啓，統計科学のフロンティア3 モデル選択，岩波書店，2004年
［6］ 徳永健伸，言語と計算5　情報検索と言語処理，東京大学出版会，1999年
［7］ 総務省，平成24年度版 情報通信白書，ぎょうせい，2012年7月，P119～121
［8］ 北研二，言語と計算4　確率的言語モデル，東京大学出版会，1999年
［9］ 林田英雄，脇森浩志，テキストマイニング技術とその応用，ユニシス技報，日本ユニシス，Vol.24 No.4 通巻84号，2005年2月
［10］ 林田英雄，Webマーケティングのための CGM分析，ユニシス技報，日本ユニシス，Vol.31 No.3 通巻110号，2011年11月

執筆者紹介　脇　森　浩　志（Hiroshi Wakimori）
2003年日本ユニシス(株)入社．テキストマイニングを始めとする情報分析・検索技術の実業務への適用に携わる．現在，Topic-Explorer，TopicStation，MiningPro21文書マイニング・システムなどテキスト分析システムの開発，適用を担当．経済産業大臣登録 中小企業診断士．