人工知能学会
知識流通ネットワーク研究会

ブログ上のクチコミ情報分析

高橋 哲朗, 岡本 青史†, 友澤 大輔††

アブストラクト: 近年，インターネットの普及により一般の人達が簡単に情報を公開することが可能となってきている．これらの情報を使うことにより，企業と消費者などの情報の伝達をより活発にすることができると考えられる．しかしこれらのデータの多くは文書で書かれているため，それらを機械的に扱うのは難しく，それらの情報を利用するためには人が 1 つずつ読まなければならないのが現状である．この課題に対して我々は，テキストから評価を表わす表現とその評価の対象となる特定の対象物の対を抽出する技術を開発し，マーケティングの分野に適用している．本稿ではこの技術のマーケティングにおける適用例を紹介し，またこの技術がブログ上における知識の蓄積と流通においてどのような役割を果たし得るかについて議論する．

Analysis of Word-of-Mouthe on Blogsphear

TAKAHASHI Tetsuro, OKAMOTO Seishi, TOMOSAWA Daisuke

Abstract: The recent innovation of the Internet allows us to publish and share various information on it. While the innovation seems to be able to actualize good communication between a municipal corporation and citizens, it has not yet due to difficulty of text processing. And people have to read them manually with labor today. In order to solve the problem, we developed a technology called 'sentiment analysis' which extracts sentiment pairs which consists of a sentiment expression and a target word such as a name of product, company, brand and so on. In this article, we first discuss how the technology can be applicable to a case of marketing. And then we show a perspective on communication of knowledge in blogsphear.

1 はじめに

近年，インターネットの普及により一般の人達が簡単に情報を公開することが可能となってきている．情報公開の場としては，掲示板やソーシャル・ネットワーキング・サービス (SNS) ，ブログなどが挙げられる．ブログとはインターネット上に公開する日記形式の記事を書くためのサービスであり，ユーザが手軽に自分のページを持つことができ，またその編集も容易にできるため特に広く普及してきている．現在その利用者は 800 万人に上り，国内だけでも 1 日に約 50 万もの記事が書かれている [4]．このような環境の基，ネットワーク上でのコミュニケーションが可能になってはきているが，そこで生み出される情報は十分に活用されていないという現状がある．その理由の 1 つとして大規模な情報の扱いの難しさが挙げられる．ネットワーク上には非常に多くの情報があるが，それらの情報は整理されていないために，必要な情報だけを引き出したり全体を俯瞰することが困難となっている．この問題に対して我々は，評判情報が人々の意見や関心を整理するための基本的な情報として役立つと考え，大量の文書から特定の製品やブランドなどに対する評判情報を自動的に抽出する技術を開発した．

†株式会社富士通研究所 (Fujitsu Laboratories LTD.)
††ニフティ株式会社 (NIFTY Corporation)

本稿ではまず我々の開発した評判情報分析技術について説明する．そして次にその適用事例の 1 つとしてマーケティングでの事例を紹介し，続いてこの技術がブログ上における知識の蓄積と流通においてどのような役割を果たし得るかについて議論する．

2 評判情報分析技術

本システムの概要図を図 1 に示す．ここに示すように，本システムではテキストから評価の対 (以降，評価対) を抽出する．抽出した評価対の集合に対してマイニング技術を適用することにより，様々な可視化を可能としている．

2.1 処理手順

本システムは以下の要素技術に分けられる．

- 記事収集・本文抽出
- 自然言語解析
- 評価対抽出
- テキストマイニング
- 可視化

2.1.1 記事収集・本文抽出

システムはまず情報源となる HTML 文書を収集する．Web 上の HTML 文書には書き手が書いた記事そのもの以外にも多くの情報が含まれているため，HTML 文書から記事のみを抽出しなければならない．

評判情報分析においてはこの処理が重要となる．たとえば一般のブログ記事においては，両サイドのフレームに，アフィリエイトを始めとするさまざまな情報が書かれている．ここには著者の書いた評判情報以外の情報が含まれるためここからの評判情報の抽出は適切ではない．

本システムでは HTML 文書の中から著者の書いた部分を推定し，その部分のみを抽出する技術を用いている．

2.1.2 自然言語解析

自然言語で記述されたテキストを計算機で処理するために，テキストを解析し構造化する必要がある．本システムでは入力されたテキストに対して以下の処理を行なう．

形態素解析

テキストを単語に分割し，それぞれの単語に品詞を付与する．

固有表現抽出・名詞句同定

組織名，製品名，人名など，評価の対象となりうる語句を特定する．機械学習に基づく手法により，辞書情報や文脈情報を考慮し上記の語句の特定を行なう．辞書にのみ依存してはいないので，新語も抽出可能である．抽出精度は未知の語に対して約 90% また，抽出のためにあらかじめ辞書に登録しておくことにより，より正確に対象物を抽出することもできる．

評価表現抽出

品詞に依存しない多様な評価表現パタンを用いて，評価表現の抽出を行なう．

これらのパタンでは形容詞以外の語も抽出対象として評価表現パタンを使っているため，より大規模に評価表現を抽出可能である．形容詞以外の表現には,「好き (動詞)」,「満足している (サ変名詞 + 動詞)」,「夢中です (名詞)」などがある．

2.1.3 評価対抽出

抽出した評価表現がどの対象物を評価しているかを見付ける必要がある．このタスクをここでは評価対抽出と呼ぶ．評価対抽出は情報抽出という技術の一つととらえられており，近年多くの研究者によって研究が行なわれている [3]．評価対抽出の方法にはいくつかの手法が考えられるが，ここでは基本的な手法を 2 つ紹介する．

まず 1 つ目は共起による抽出である．この手法では，対象とする商品やサービスの書かれている記事内に評価表現が出現していた場合にそれらを評価対として抽出する．この手法では正しい評価も抽出できるが，正しくない評価まで大量に抽出する点が問題点として挙げられる．たとえば 図 1 の例では，“NX →重い”，“LOOX →気に入って” という正しい評価だけでなく，“LOOX →重い”，“NX →気に入って” などの誤った評価まで抽出してしまう．つまりこの手法は，カバレッジは高いが精度が低いと言える．予備調査の結果，この手法を使った場合の精度は約 3% であるという結果を得た．

もう 1 つの手法として，単語間の文法的な係り受け情報を用いて評判情報の抽出を行なうものがある．この手法は，精度は高いがカバレッジが低いと言える．たとえば 図 1 の文章では，“NX” と “重い” の間には係り受けの関係があるので，これらを評価対として抽出可能であるが，“NX” と “NG”，また “LOOX” と “気に入って” の表現は直接の係り受け関係にないので，これらを評価対として抽出できない．特に日本語の文書においては，一度出現した語は区別の必要がある場合を除いて省略される傾向にあるのでこの問題が顕著になる．係り受けのみを用いた場合に抽出できない情報の割合を予備調査した結果，約 80% を抽出できないと

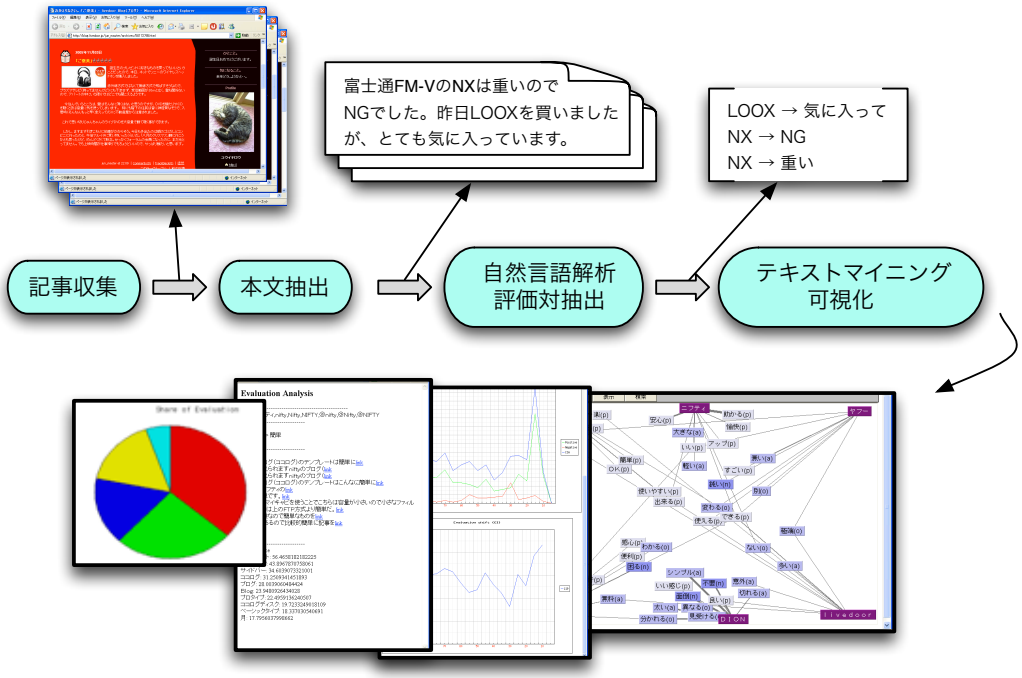

図 1: 評判情報分析システムの処理概要

いうことが分かった．

上記の問題を解決するために我々は，

- 機械学習に基づく手法を用いて評価対となりうる候補を選択し，
- そこから頻度に基づくテキストマイニグを行なう．

という手法を提案し実装した．

評価対抽出では，まず評価を受ける対象物と評価表現の組み合わせ対を “LOOX →良い” のような形で列挙し，これらを評価対の候補とする．次に学習器と分類器を用いてこの候補の中から正しい関係にある評価対を分類し，選択された評価対から頻度に基づく評価対のマイニングを行なう．機械学習では，数千記事に対して人手で評価対の情報を付与し，それを元にどのような文脈のときに評価対の候補が評価対となるかを学習させた．この手法を用いることにより，カバレッジを高く保ったまま精度を上げることを実現している．

2.1.4 テキストマイニング

評価対の出現頻度によるマイニングを行なうことにより，精度を高めることができる．概念的には，より多数のブログで語られている意見はより確からしいという仮定に基づき信頼度を設定し，その信頼度の高い評価だけを用いる．評判をユーザに提示する場合を考えると，ユーザが見られる記事数は限られているので，上位 N 位 (たとえば N=10) だけを用いることができ，出力における精度を上げることができる．

表 1: 評価対

ブログ URL	対象物	評価表現
http://AAA.xxxxx.html	FM-V	使いやすい
http://BBB.xxxxx.html	LOOX	軽い
http://CCC.xxxxx.html	LOOX	小さい
:	:	:

また，ブログの持つ特徴の 1 つに，情報の発信者を擬似的に特定できるという点が挙げられる．個人を特定することはできないが，ブログの URL を個人の ID と見なすことにより，ある評価を書いた人が過去にどのような評価を書いていたか，また他の製品についてはどのような評価を書いているか，などを知ることができる．このデータを 表 1 のような形で整理することにより，POS[3]データに対する集計と同じような集計が可能となる．

3 マーケティングへの技術適用

2 節で説明したように，本システムは特定の製品・ブランド・企業などについての評判 (好き，欲しい，かわいい等) を抽出することができる．この技術はすでに BuzzPulse[4]というサービスにおい使われている．本節では，BuzzPulse で行なっているマーケティングへの適用について述べる．

[3]Point of Sales: 店舗における販売実績
[4]http://www.nifty.com/buzz/：ニフティ株式会社

本稿で紹介した技術を用いることにより特定の製品やブランド，企業についての評判情報を得ることができ，この評判情報は広告やキャンペーンの効果測定，また新製品開発などに用いることができる．

たとえばある企業が新製品の発売に際し新しいテレビCMを始めたとき，そのCMによりインターネット上の評判の量がどのように変化したかを見ることにより，訴求した内容がどれくらい消費者に届いているかを知ることができる．評判量が増加すればそのキャンペーンは成功だったと言え，変化しなければ効果が無かったと言える．このようにして得られた事例を集めることにより，どのような製品にはどのようなキャンペーンをどの媒体によりどのように行なうべきかといった分析が可能になる．従来，テレビにおける視聴率やインターネット上の閲覧数などの数値を測定することは可能だったが，その結果人々がどのような感想を持ったかまでは測定できなかった．本稿で紹介した技術はそれを可能にするものである．

また単純な評判の数だけでなく評判の内容も考慮することにより，より多くの知見を得ることができる．たとえば新製品についての評判の内容を見たときに機能面に関する評判が多く書かれていることが分かったが，それと同時にデザインについての評判が少なかった，ということが分かったとすると，その製品はデザインについてより改善する必要があるといった新製品開発に有用な知見をそこから得ることが可能となる．

ブログはそれぞれ固有の URL を持っているため，この URL により著者を区別することができる．そして評価対を用いることにより，車好きな著者やお菓子好きな著者をあらかじめ識別できる．単純に評判の量を得るだけではなくこの著者の種類も使うことにより，たとえば「今回の新製品は車好きにはうけたが，音楽好きの人には評判が良くないようだ」といったよりターゲットをしぼったマーケティングや分析が可能となる．

4 蓄積される知識とその活用

4.1 ブログ上にある知識

3 節で紹介したマーケティングの例において，企業の求める主な情報は特定の商品の評判であるが，ブログの中に評判が直接書かれている訳ではない．ブログにはブログ著者の個別の意見や経験が書かれている．そのため特定の商品に対する世の中全体からの評判を知るためには，それらの個別の意見や経験を抽出し，整理する必要がある．

個別の意見や経験の抽出においても困難な点がある．知りたい情報は消費者の意見や経験であるが，ブログにはそれらがそのままの形で書かれているのではなく，また整理された形でまとめられているのでもない．ブログには，個別の消費者が持った意見や経験がその消費者の言葉により文書の形に記号化 (encode) された情報として書かれている．そして我々の扱える情報はその記号化された情報のみであるため，この情報を扱うためには復号化 (decode) する作業が必要となる．

たとえば，例 (1a) では「おどろきでした」という表現が良いこととして書かれているのか悪いこととして書かれているのかは文脈に依存する．この文の後に,「こんなに使いやすいんですね。」と書かれていれば良い評価をしていると言えるが,「起動に時間がかかりすぎです。」と書かれていれば悪い評価をしていると言える．

(1) a. 富士通の新しいパソコンを使ってみましたが、おどろきでした。

b. この車はハンドリングがずいぶん固いね。

例 (1b) は，筆者が「ハンドリングが固い」ということを好むか好まないかによって評価の極性が異なり，この文脈からだけでは正しく推測することができない．これは著者依存の問題と言うことができ，この問題を解くためにはユーザモデリングの概念が必要となる．

ブログを用いた評判情報分析には，ブロブ上に書かれていることがすべてではないという問題もある．たとえばある製品について「使いやすい」という評価が書かれていた場合，その記事の著者はその製品について「使いやすい」という評価をしたことが推測できる．しかし，たとえば「重い」という評価が書かれていなかった場合に，それは「軽い」ということが含意されているのではなく，また重さについての評価が行なわれたのかどうかも分からない．

つまりブログ上の情報から人の意見や経験を得ようとする場合，部分的な情報しか取ることができないことに注意する必要がある．たとえば，商品 A と商品 B を比較したとき，ブログ上では商品 A の方がポジティブな意見をより多く得ていたとしても，世の中の評判がそれだけ A を支持していると言うことはできない．商品 B の方が世の中から良い評判を得てはいたが，

- たまたまブログ著者への認知度が低かった
- たまたまブログには書かれなかった
- 良い評価が，直接的な表現では書かれなかった

などの可能性が残る．

これらの問題に対する本質的な解を求めることは非常に困難である．そのため我々は現在，著者の持っている意見や経験とテキスト中での記述は同一であるという仮定を置き，2 節で述べた手法で抽出された評価対

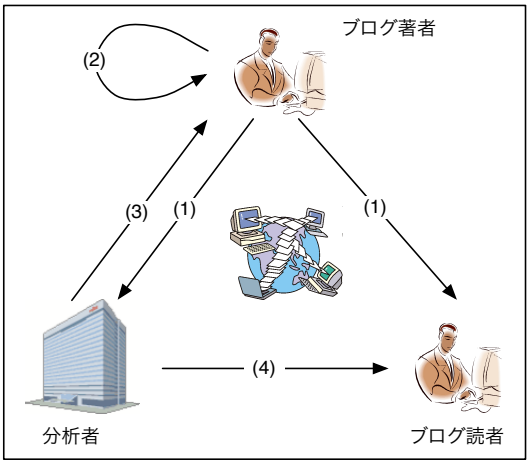

図 2: 知識の流れ

を著者の評価と見なしている．そこでは大規模な表現のパタンセットを構築しており，できるだけ網羅的に評価を表わす表現の抽出を試みている．しかし，パタンを用いるため文脈非依存非依存でしか抽出できないという問題は残る．またブログ著者への認知度や，ブログに書かれる確率なども比較対象間で均等であるという仮定を置いて分析を行なっている．厳密にはこれらを加味して分析を行なうべきであるが，これらの情報はブログのみから得ることはできない．

4.2 知識流通における効果・役割

本節では評判情報分析の果たす役割を，知識の流れという観点で整理する．ドメインとしては，3 節で紹介したマーケティング分野に焦点を当てる．またインターネットを流れる情報の中でマーケティングに有用なものとして消費者の属性情報や購買情報などがあるが，これらは一般的にオープンになっていないためここでは対象外とし，ほぼ自由にアクセス可能なブログの情報から取り出すことのできる知識に焦点を当て議論する．

知識の流れは 図 2 のように図示できる．この流れの要素としては，ブログの筆者と読者，そして評判を分析する分析者がある．マーケティングにおいては，著者と筆者は消費者であり，分析者は企業ととらえることもできる．この流れの中で，蓄積された知識がブログを媒体として移動する．

ここで技術の果たし得る効果や役割としては，知識の可視化とその流れの活性化が挙げられる．

4.2.1 知識の可視化

流れる情報には一次的なものとして個別の意見・経験がある.「一次的」というのは，これらの情報源である筆者から発信された未加工の情報という意味である．これらの未加工の情報は，そのままでも読むことはできる．つまりブログの記事に書かれた個別の意見や経験は，その記事を読めば得ることのできる情報である．しかし，ブログ記事の量は一般的に大量であるため，それら全体の内容を把握するためには非常にコストがかかるという問題がある．商品の購入を検討している消費者が特定の製品について調べるときや，企業のマーケターが自社製品の評判の調査を行なうときに，その製品の名前が表われるブログ記事をすべて読むことは現実的ではない．

本稿で紹介した評判情報分析技術は，この情報の整理や可視化を可能にするものであると言える．この技術は 2 節で紹介したように，テキスト情報から表 1 に示すような構造化された情報を抽出する．一度このような構造化された形にできれば，

- ある製品が世の中 (ブログ全体の中) で得ている良い評価と悪い評価の割り合い
- 多くの人から言われている評価 (e.g.「使いやすい」など)
- 他商品と比較したときの，評判の量や質 (ポジティブな評価の割合いなど)

といった形で，容易に整理し可視化することが可能となる．

4.2.2 知識流通の活性化

4.2.1 節で述べた「可視化」により情報の流通がより活性化されることが期待できるが，より積極的に「活性化」を行なうアプローチも考えられる．

本稿で紹介した技術は，テキスト中の個別の意見や経験を抽出するというものであった．つまり情報の流れの中では，図 2 の (1) の部分をサポートしていると言える．個別の著者の書いた評価から，その著者の持つ意見や経験に関する知識を推測し，それらを集約することにより分析者やブログ読者は評判情報を得ることができる．

これに加えさらなる活性化の方法として以下のような項目が考えられる．

著者→著者 (2)

ブログの著者がある記事を書いているときにその内容に関連する情報を提供することにより，そこでの記述の内容をより深いものにすることが可能となる．たとえば，著者がある製品 A について「使いやすい」という意見を書いているときに，同じ意見が書かれた記事を提示したり，または逆の「使いにくい」という意見が書かれた記事を提示することにより，その著者に

より多くの情報を与えることができる．言い換えると，著者の知識のみを用いて書くだけではなく，他者の知識も同時に参照しながら，その上での新しい知識を生み出すことを可能とすることになる．

企業→読者 (3)

企業は独自の情報を各々のホームページ等で配信しているが，多くの消費者は，商品の購入に際し企業の情報だけでなく他の消費者のクチコミを頼りにしている．そこで企業としても一般消費者のクチコミを収集しその中の良い評判を掲載することにより，より大きな広告効果を期待できる．また良い評判だけでなく悪い評判や不満などについても，それらに対する回答や対応をアピールすることにより良いイメージを構築するなどの戦略が考えられる．

企業→著者 (4)

企業が特定のブログ著者に対してコンタクトを取り，新しい情報や新商品などを提供することにより，それに関しての記事をブログに書いてもらうという，バズマーケティングやインフルエンサーマーケティングと呼ばれるマーケティング手法がある．このような手法に対しては記事の公平性に関する是非はあるが，ブログへの記述が強制的でない場合は，テレビ CM や雑誌，街頭でのサンプル配布などの認知経路の一つとしてとらえることができ，企業とブログ著者が Win-Win の関係を築ける可能性がある．

この手法においては，企業にとってよいブログ著者をいかにして見付けるかが重要な鍵となる．企業から提供される情報や商品などに対して共感してくれる人を見付けなければ，その効果が得られにくいと考えられる．本稿で紹介した評判情報分析技術を用いることにより，たとえば良い意見を多く書いている著者ほど好ましいなどの指標を用いて，よいブログ著者を見付ける手助けができると考えている．

4.3 課題

ブログ上の知識流通における課題としては信頼性がある．この信頼性には少なくとも 2 つのレベルがあり，その 1 つは自動的に生成されるスパムブログである．このタイプのスパムブログは現在増加してきており，これらを適切に排除できなければ本稿で述べたような知識の流れを確立することはできない．

2 つ目のレベルの信頼性の課題は書き手の多様性によるものである．本稿で対象としているブログは Web 上のさまざまな人により書かれているため，その内容や質は玉石混淆である．そのため，ブログそのものやブログから得られる情報に信頼性があるかどうかという議論がある．確かにブログの情報がすべて信頼できる訳ではないが，アンケート調査等においてもある一定量の正しくない情報は入り得てしまい，ブログにおける虚偽の割り合いと比較することは非常に困難である．

ブログを書く動機という点に着目すると，一般的なアンケート調査では対価が支払われることが多く，その対価を目的として答えられる可能性があるが，それに対してブログは，対価なく書かれるという点が異なる．社会心理学においては，ブログを書く動機は自分のための備忘録や日誌として書かれる場合が多いという調査結果が山下ら [5] から示されている．また Cornwell ら [1] は，対面での対話とコンピュータを介した対話における嘘の比率には差はなかったという調査結果を示しているおり，Joinson [2] の実験では，コンピュータを介した議論の方が，対面での議論と比べ自己開示の度合いが高かったことを示している．これらの研究から，アンケート調査などと比較したときにブログの方が調査信頼性に欠けるとは一概には言えない．

5 まとめ

本稿では評判分析技術を紹介し，その適用分野として，マーケティングを挙げこの分野でどのような利用が可能かを議論した．

インターネットの発達により，情報の蓄積や通信そのものの技術は飛躍的に進歩し多くの情報が行き交うようになったが，その量の多さのために必要な情報を的確に得ることが難しくなってきており情報を活かせないでいる．そのような中で，本稿で紹介した技術を用いることにより的確な情報にアクセスする枠組を用意できれば，より効率的・効果的な情報の伝達を可能にすることができるようになる．

参 考 文 献

[1] B. Cornwell and D.C. Lundgren. Love on internet: Involvement and misrepresentation in romantic relationships in cyberspace vs. realspace. *Computers in Human Behavior*, Vol. 17, pp. 197–211, 2001.

[2] A.N. Joinson. Self-disclosure in computer-mediated communicaton: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, Vol. 31, pp. 177–192, 2003.

[3] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, 2006.

[4] 財団法人インターネット協会. インターネット白書 2006. 2006.

[5] 山下清美, 三浦麻子. 人はなぜウェブ日記・ウェブログを書き続けるのか (2). 日本社会心理学会第 45 回大会論文集, pp. 694–695, 2004.