センチメント分析とトピック抽出によるマイクロブログからの評判傾向抽出

センチメント分析とトピック抽出によるマイクロブログからの評判傾向抽出

橋本　和幸† 　中川　博之† 　田原　康之† 　大須賀昭彦†

Reputation Trend Extraction from Microblogging Using Sentiment Analysis and Topic Extraction

Kazuyuki HASHIMOTO†, Hiroyuki NAKAGAWA†, Yasuyuki TAHARA†, and Akihiko OHSUGA†

あらまし　近年，政策やサービスなどの評判を調査するためにアンケート形式による調査が増加しているが，回収率が低落傾向にあることや，人的コストが増加するなどの問題が生じている．一方，Web 上にはユーザの意見を含む評価情報が多数存在している．そこで本研究では，評判傾向予測システムの構築を目的として評判傾向の時間的変化とその原因をマイクロブログから抽出する評判傾向抽出エージェントの実現を目指す．特に本論文では，評判傾向の抽出と評判傾向が変化した原因の抽出に注力する．評価情報の感情を抽出するセンチメント分析に着目し，回帰式から評判傾向の変化点を抽出した後に，変化点におけるトピックをチャンキングにより抽出する手法を提案する．本手法は，従来の評価判定法である p/n 判定にセンチメント分析を組み合わせることで，p/n 判定単体よりも人手による調査と相関の高い時系列変化を抽出できる点が特徴であり，政治及びテレビドラマに関するコンテンツを対象に実際の支持率，視聴率に対する評価実験を実施した結果，政治（自由度調整済決定係数 R'^2（p/n 判定単体：0.22，提案手法：0.60)），テレビドラマ（自由度調整済決定係数 R'^2（p/n 判定単体：0.26，提案手法：0.56））ともに相関の高い時系列変化を抽出できることが確認できた．

キーワード　マイクロブログ，評判抽出，センチメント分析，トピック抽出，Web エージェント

1. まえがき

近年，政策やサービスの評価や評判の傾向を調査するために，アンケート形式による世論調査や市場調査の頻度が飛躍的に増加している [1]．しかしながら，これらの調査の回収率は低落傾向にあり，人的コストや時間的コストが増加するとともに，顧客のニーズの変化が激しいことから短期的な調査が必要となっている．また，継続的な調査が困難なため，調査結果に何が影響を与えたのかといった原因の抽出が難しい．その一方で，ブログやソーシャル・ネットワーキング・サービス（以下，SNS）などの Web サービスの普及により，誰でも簡単に情報を発信できるようになってきている [2]．これらのユーザコンテンツには政策，製品やサービスに関する意見などの評価情報を含むものが多数存在している．

そこで本研究では，このような Web 上のユーザコンテンツから評判傾向を予測するシステムの構築を目的として，評判傾向の時間的変化と，その原因を抽出する評判傾向抽出エージェントの実現を目指す．本研究により，世論調査や市場調査のコストの削減や，顧客のニーズの時間的変化を把握することができるようになると考えられる．

特に本論文では，評判傾向の抽出と評判傾向が変化した原因の抽出に注力する．まず，リアルタイム性の高い評判傾向を抽出するために，マイクロブログを評判傾向抽出のための分析対象として用いる．マイクロブログとは，ユーザが短いテキストを書き，不特定多数，若しくは限定されたユーザに公開する形態のブログである [3]．次に評判傾向の抽出には，評価情報を抽出するために評価情報を肯定的/否定的に分ける p/n 判定 [6] と，評価情報の感情を抽出するセンチメント

† 電気通信大学大学院情報システム学研究科，調布市
Graduate School of Information Systems, The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan

分析を利用する．また，変化の原因をより明確化するために，形態素間の出現確率に着目したチャンキング手法によるトピック抽出を試みる．チャンキング手法により形態素単位の重要語ではなく，連続する複数の単語により構成されるトピックを抽出することが可能となる．

更に本論文では，本手法の有効性を検証するために政治，及びテレビドラマに関するコンテンツに対して本手法を適用した実験結果を示す．本実験結果から，政治（自由度調整済決定係数 R'^2（p/n 判定単体：0.22，提案手法：0.60)），テレビドラマ（自由度調整済決定係数 R'^2（p/n 判定単体：0.26，提案手法：0.56））の結果を得ることができた．これにより従来の評価判定法である p/n 判定にセンチメント分析を組み合わせることで p/n 判定単体よりも人手による調査と相関の高い時系列変化を抽出できることが確認できた．また本実験結果から，評判傾向の変化の原因となったと考えられるトピックが抽出されることも確認できた．

以降，本論文は次のような構成をとる．まず **2.** で評判傾向抽出の課題，**3.** で本研究のアプローチについて述べ，**4.** でマイクロブログの構造と特性に着目した提案手法を述べる．**5.** では，本手法の有効性を検証するために実施した評価実験と考察を記載し，**6.** で関連研究について述べる．最後に，**7.** で今後の課題と合わせて本論文をまとめる．

2. 評判傾向抽出の課題

本研究では，Web 上のコンテンツから評判傾向を予測するシステムの構築を目的として，評判傾向の時間的変化と，その原因を抽出する評判傾向抽出エージェントの実現を目指す．これまでにも評判傾向の時間的変化や，変化の原因抽出に関する研究がなされている [4] が，解決の必要がある課題が幾つかある．

2.1 課題 1：評価情報の抽出

まず，口コミサイトの投稿やブログを解析し評判を抽出する研究 [5] では，扱っているコンテンツの文字数が比較的多く，ユーザの意見が婉曲的な表現になりやすいため，一般に評価情報の抽出が難しい．また，多くの研究では評価情報を肯定的/否定的（肯定極性/否定極性）の二つの評価極性に当てはめる p/n 判定を用いてコンテンツから評価情報を抽出している．しかし，肯定極性及び否定極性のいずれにも該当しないコンテンツも数多く存在する [4]．

- この政策には驚いた
- この音楽は懐かしい

例えば，上記に示すように “驚く” や “懐かしい” などの語は評価を含むものの明示的な評価ではないため，肯定極性/否定極性に当てはめることができたない．このように p/n 判定では抽出することができない評価情報が存在する．

2.2 課題 2：評判傾向の変化の抽出

また，評判傾向の変化を抽出するためには，大量のコンテンツから評価情報を抽出する必要があり，それらのコンテンツのリアルタイム性が高いことが望ましい．これまでの評判傾向の変化を抽出する研究 [12] では一般的にブログを解析対象にしている場合が多い．しかし，ブログの実態に関する調査研究の結果によると，日本国内で 1 か月に 1 回以上記事が更新されているブログは全体の 2 割であるとされている [8]．よって，全体の 8 割が 1 か月以上更新されていないため，ブログから 1 日単位や 1 時間単位の評判傾向の変化を抽出することは困難である．

2.3 課題 3：変化の原因抽出

加えて，変化の原因抽出の研究 [7] では，評判傾向が変化した時間における形態素単位の重要語抽出が一般的であり，タグクラウドなどの出力が多い．しかしながら，タグクラウドは重要な単語群を表示しているため，この出力結果から変化した原因を推測することが必要であり，明示的な原因が分かりにくい．

3. 本研究のアプローチ

本章では，**2.** で述べた課題を解決するための本研究のアプローチについて述べる．

3.1 アプローチ 1：センチメント分析

まず本研究では，課題 1 の「評価情報の抽出」を解決するために，従来手法である p/n 判定とともに，コンテンツの感情を抽出するセンチメント分析を実施する．センチメント分析とは，コンテンツのセンチメントを喜（喜ぶ）・怒（怒る）・哀（悲しむ）・怖（怖がる）・恥（恥じる）・好（好む）・厭（厭がる）・昂（昂ぶる）・安（安らぐ）・驚（驚く）の 10 種類の感情表現に分類する手法である．このような感情を抽出するセンチメント分析を用いることで p/n 判定では抽出することのできない評価情報の抽出を目指す．

3.2 アプローチ 2：マイクロブログの利用と重回帰分析による調査対象のモデル化，評価表現の特定

また本研究では，課題 2 の「評判傾向の時間的変化」

を解決するためにマイクロブログの代表的なサービスの一つである twitter [9] に着目する．twitter は特徴としてコンテンツに 140 文字の文字数制限があり，気軽に投稿できるため，ブログや，SNS などと比べてユーザコンテンツのリアルタイム性が高く，コンテンツ数が多いといわれている [10], [11]．このような性質をもつ twitter を用いることで，詳細な評判傾向の時間的変化の抽出が期待できる．

また本研究では，調査対象の評判傾向のモデル化と調査対象に影響を与える評価極性，感情表現を特定するために重回帰分析を実施し，特定した評価表現（評価極性及び感情表現）の出現推移を元にした評判傾向の時系列変化を抽出する回帰式を構成する．その後，構成した回帰式から評判傾向が変化した変化点を抽出する．このように重回帰分析を実施することで，調査対象の評判傾向のモデル化と調査対象（支持率や視聴率の変化など）ごとに着目すべき評価表現を特定することができる．

3.3 アプローチ 3：チャンキングの利用

更に，課題 3 の「変化の原因抽出」を解決するために，得られた回帰式の変化点におけるトピックを抽出する．このトピック抽出には，形態素間の連続する確率に着目するチャンキング手法を用いる．チャンキングとは，主に文中の語の関係性を同定する手法である．この手法は，形態素単位の重要語ではなく，連続する複数の単語からなるトピックを抽出するものである．

4. 提案手法

本研究では，評判傾向の時間的変化と，その原因を抽出する評判傾向抽出エージェントを提案する．図 1 に提案手法の概要を示す．図 1 のようにエージェントは，twitter 上の特定のキーワードに対するコンテンツを収集し，twitter の特性を生かした p/n 判定とセンチメント分析を実施することで評価情報（評価極性値あるいは感情表現値）を抽出する．本研究では，コンテンツ中の評価表現に基づいてコンテンツを p・n に決定したものを評価極性値とし，同様に喜・怒・哀・怖・恥・好・厭・昂・安・驚のいずれかに決定したものを感情表現値とする．次に，重回帰分析を用いて調査対象の評判傾向のモデル化と調査対象に影響を与える評価表現を特定し，回帰式を構成する係数を与えることで，エージェントは回帰式を用いて評判傾向の変化点を抽出し，その変化点におけるトピックを出力する．以降では，twitter の特徴を生かした評価情報の抽出，重回帰分析による調査対象のモデル化と評価表現の特定，変化点の抽出，変化点でのチャンキングを利用したトピック抽出手法を順に述べる．

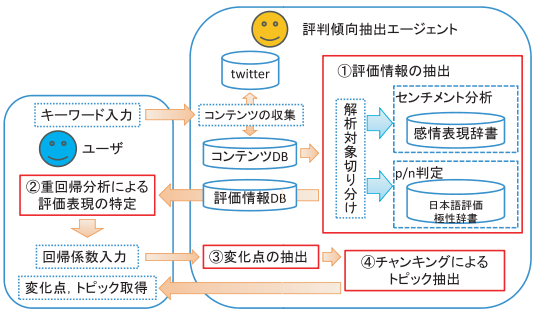


図 1 提案手法の概要
Fig. 1 Overview of proposed agent.

4.1 評価情報の抽出

ユーザコンテンツの評価情報を抽出するために，本研究では p/n 判定とセンチメント分析を実施する．p/n 判定とは，評価情報を肯定的/否定的に分ける手法であり，本論文では日本語評価極性辞書 [14] を用いる．一方センチメント分析とは，ユーザコンテンツの心的態度を抽出する手法であり，本研究では感情表現辞典 [13] を参考に感情表現辞書を構築する．感情表現辞典は，喜・怒・哀・怖・恥・好・厭・昂・安・驚の 10 種類の日本語表現，単語を収録したものである．この辞書中に含まれる旧字体（目出度い，忌忌しい）を新字体（めでたい，忌々しい）に変換したものを追加し，感情表現辞書とする．本手法では両辞書を使用し，ユーザコンテンツから特定のキーワードに対する評価情報を係り受け解析器 CaboCha [15] と形態素解析器 MeCab [16] を用いて抽出する．

ここで，エージェントがユーザコンテンツから投稿者の評価情報を抽出するために，コンテンツ中で投稿者の意見が現れている部分を解析対象にする必要がある．本論文では，コンテンツ中の投稿者の意見が反映されている部分を twitter の機能と特性に着目することで特定する．twitter には他のユーザの投稿を転送する *ReTweet*（*RT@UserName*），引用する *QuoteTweet*（*QT@UserName*），明示的に特定のユーザに対して投稿する *Reply*（*@UserName*）の三つの機能（以下，*RT*，*QT*．@）がある．ここで本論文では投稿者自身の意見を抽出するために，コンテンツの先頭から最も近い機能を示す文字列の直前の文字列を「機能直前の文字列」，コンテンツの先頭から

最も近い機能を示す文字列の直後の文字列を「機能直後の文字列」とする．例えばコンテンツが「AAA RT @UserNameA BBB QT @UserNameB CCC」とした場合，機能直前の文字列は「AAA」であり，機能直後の文字列は「BBB」である．「AAA」は投稿者が作成（追記）した部分であり，「BBB」に対する意見であることが多いため，本論文では解析対象の候補としている．加えて，このように機能直前の文字列を解析対象とすることで，「CCC」のような投稿者の意見が反映されにくい文字列を解析対象外とすることが可能である．また，「RT @UserNameA BBB QT @UserNameB CCC」の場合，機能直後の文字列である「BBB」を解析対象の候補としている．これは，「BBB」は投稿者が転送した記述であり，投稿者の意見が示されることが多いと考えられるからである．このように投稿者の意見が示されている部分に着目することで評価表現を抽出する．図 2 のように，キーワードを “〇〇党” とすると，投稿者は初めのコンテンツでは *RT* 直前，二つ目では *RT* 直後にそれぞれ自分の意見を述べている．このように投稿者の意見となり得る場所を解析対象とすることで，引用文に含まれる評価表現を排除した，投稿者自身の評価を抽出することができる．よって，機能有無，及び特定のキーワードの位置と機能の直前または直後の文字列に着目して場合分けし，解析対象を決定する．図 3 に，この特性を生かした解析対象の切分けを示す．図 3 中の 1，2 の場合は係り受け解析器 CaboCha を用いてキーワードが直接係る文字列と辞書をマッチングし，評価情報を抽出する．例えば，図 4 の (1) ではキーワードが “〇〇党” であることから，解析対象は “〇〇党は嫌い。” となり，キーワードが係る文字列は “嫌い” となる．図 3 中の 3 の場合は形態素解析器 MeCab を用いて形態素解析後，各文字列の形態素と辞書をマッチングし，評価情報を抽出する．このようなコンテンツは，引用部分のコンテンツに対しての意見を含む場合が多いため，機能直前の文字列を解析の対象とする．例えば図 4 の (3) ではキーワードが “〇〇党” であり，キーワードが転送の機能を表する *RT* 直後にあることから，キーワード直前の文字列 “嫌い。” を形態素解析し，辞書とマッチングする．図 3 中の 4 は機能を使用していないコンテンツが該当し，これらは全文字列を解析対象とし，係り受け解析器 CaboCha を用いてキーワードが直接係る文字列と辞書をマッチングする．その後，解析対象中に含まれる評価表現数を数え上げ，最後に評価表現数が最大の評価表現をそのコンテンツの評価情報とする．例えば，抽出した評価表現が，“厭”2 回，“怒”1 回の場合，そのコンテンツの評価情報は “厭” とする．

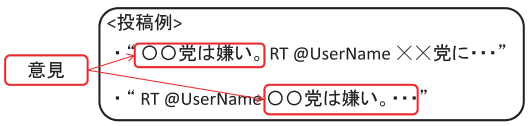


図 2　コンテンツ中の意見例
Fig. 2　Ex. opinion in microblogging.

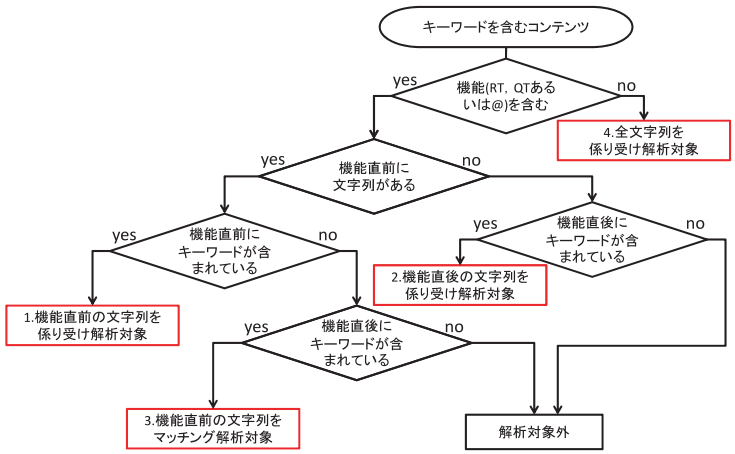


図 3　twitter の特性を生かした解析対象の切分け
Fig. 3　Analysis target distribution based on twitter’s functions.

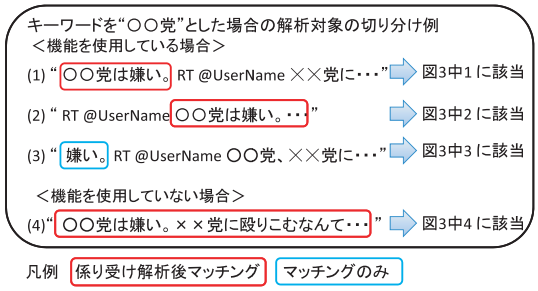


図 4　解析対象の切分け例
Fig. 4　Examples of analysis target distribution.

4.2　重回帰分析による調査対象のモデル化と評価表現の特定

以上のようにしてユーザコンテンツから評価情報を抽出し，調査対象の評判傾向のモデル化と調査対象に影響を与える評価表現を特定する．重回帰分析は，一つの目的変数を，どの説明変数がどの程度影響を与えているかを調査する統計手法である．式 (1) は，回帰式を得るために用いる目的関数と説明変数の関係式である．

$$Y_t = \alpha + \sum_{i}^{n} \beta_i X_{[t-1,t]i} \qquad (1)$$

本手法では，式 (1) に示すように，目的変数として調査日 t における実際の調査データ（例えば支持率など）の増減度 Y_t，説明変数に実際の調査データの調査日間 $[t-1,t]$ における n 個の評価表現の出現頻度の平均の増減度 $X_{[t-1,t]}$ を入力として与え，結果として，定数項 α と各説明変数の回帰係数 β_i を決定し，それぞれ α'，β_i' とする．得られた結果に変数減少法を適用し，調査対象に影響を与えない評価情表現を削除することで調査対象の増減に関する回帰式（式 (2)）を得る．

$$Y_d = \alpha' + \sum_{i}^{n'} \beta_i' X_{[d-1,d]i} \qquad (2)$$

4.3　変化点の抽出

得られた回帰式を用いて評判傾向が変化した時間を特定する．回帰式に日ごとの調査対象に影響を与える評価表現の増減値 $X_{[d-1,d]}$ を入力し，日ごとの調査対象の変化値 Y_d を得る．変化値が大きい場合，ユーザの評判傾向が変化したと仮定し，その日を変化日とする．その後，変化日における変化値が大きい時間を評判傾向の変化点とする．

表 1　コンテンツ数と引用を含むコンテンツ数との相関値
Table 1　A correlative value between the number of the contents and the number of the contents including the quotation.

菅内閣	民主党	自民党	公明党
0.93	0.90	0.86	0.91
共産党	国民新党	新党改革	社民党
0.80	0.75	0.75	0.81
たちあがれ日本	みんなの党	幸福実現党	無所属
0.73	0.93	0.87	0.86

4.4　チャンキングを利用したトピック抽出

3.3 で同定された変化点において発生した変化の原因（トピック）を抽出するために，本研究ではチャンキング [17] を用いる．チャンキングとは，単語を形態素単位ではなく，連続する複数の単語を抽出する手法である．ある時間における出現頻度が多い文字列を連結し，その時間における最も出現した文字列を抽出する手法である．形態素単位での重要語をコンテンツから抽出する手法が多く用いられているが，本研究では，重要語ではなく，トピックを抽出することを目的とするためにチャンキングを用いる．

筆者らは事前にチャンキングの有効性を確認するために，日ごとの特定のキーワードを含むコンテンツ数と特定のキーワードを含み引用したコンテンツ（RT（QT）を含むコンテンツ）数の関係を調査した．内閣名，政党名をキーワードとし，2010 年 6 月 24 日から同年 8 月 9 日までのコンテンツ 594,722 件を調査対象とした．表 1 は，コンテンツ数と引用を含むコンテンツ数との相関値を示している．表 1 に示すように，管内閣 (0.93)，民主党 (0.90) など比較的高い数値であり，特定のキーワードを含むコンテンツの数と，特定のキーワードを含み引用したコンテンツ数には強い相関があることが分かった．つまり，特定のキーワードを含むコンテンツ数が増えるとき，特定のキーワードを含み引用されたコンテンツ数も増加するため，形態素間の連続する確率に着目するチャンキングによるトピック抽出はマイクロブログのようなサービスに適用しやすいと考えられる．

本論文で用いるチャンキングを利用したトピック抽出手法を Algorithm 1 に示す．変化点におけるコンテンツ集合 C とし，$c_i \in C$ を形態素解析した形態素のリスト集合 $T = \{t_0, \ldots, t_m\}$ とする．起点となる形態素を t_{begin}，着目する形態素を t_{target} とし，連続する形態素間の条件付き確率を求めることでチャンクスコアを算出する．チャンクスコアとは，形態素間の関連度

Algorithm 1 チャンキングを利用したトピック抽出

入力 :変化点におけるコンテンツ集合 C
出力 :チャンキングしたトピック候補 $chunkWords$

$c_i \in C$ を形態素解析した形態素集合 $T = \{t_0, ..., t_m\}$ とする
$begin = 0$, $target = 1$
repeat
  $chunkScore = 1, n = 1$
  $target = chunkMath(chunkScore, target, n)$
  $end = target - 1$
  if $begin \neq end$ **then**
    $chunkWord = \langle t_{begin,...,end} \rangle$
    $chunkWords = chunkWords \cup \{chunkWord\}$
  end if
  $begin = end$
until $begin \leq m$

function $chunkMath(chunkScore, j, n)$
$chunkScore = chunkScore \times P(t_{j-1}|t_j)P(t_{j+1}|t_j)$
if $chunkScore > threshold^n$ **then**
  $j = j + 1, n = n + 1$
  $j = chunkMath(chunkScore, j, n)$
  return j
end if

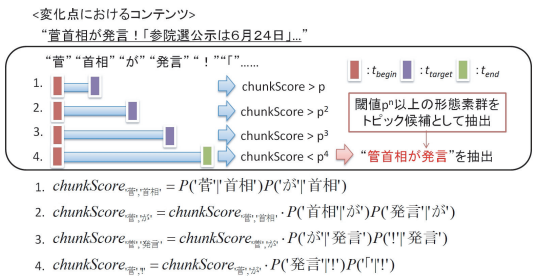


図 5　チャンキングによるトピック抽出例
Fig. 5　Topic extraction using chunking.

を示すものである．チャンクスコアが高くなるのは計算対象となっている形態素群がその時間において多く連続して存在する場合である．チャンクスコアがしきい値以上であれば，$j = j + 1$ とし，しきい値以下になるまで繰り返す．その後，終点の形態素を $t_{end} = t_{j-1}$ とし，t_{begin} から $t_{end}(\neq t_{begin})$ までの連続する文字列をチャンクワードとする．以降，$t_{begin} = t_{end}$ とし，t_{begin} がそのコンテンツの最後の形態素 t_m になるまで実行する．すべての $c_i \in C$ に対して以上を実行し，チャンクワードとして抽出した文字列をトピック候補とし，「だ.」「か.」などの句読点と付属語のみのトピック候補は除外する．変化点における出現回数が多い順にトピック候補をランキングし，トピックを抽出する．

例えば図 5 のように，変化点におけるコンテンツを形態素解析し，先頭の形態素と 2 番目の形態素からチャンクスコアを求めていく．この例の場合,「菅」「首相」「が」「発言」は変化点のコンテンツ群中で多く連続しているため，“菅首相が発言” がトピックとして抽出される．

5. 評価実験

本章では，本手法の有効性を評価するために，政治とテレビドラマに関するコンテンツに対して提案手法を適用した評価実験結果について述べる．まず，本手法によるコンテンツの切り分けの結果を調査し，解析対象の切分けが適切であったかを検証した（実験 1）また，本手法とマスメディアが定期的に調査する世論調査，及びテレビドラマ視聴率の時間的変化と類似しているかどうかを検証した（実験 2）．この検証にあたっては，p/n 判定，センチメント分析それぞれ単体で重回帰分析をした場合の結果と，両者を併用した場合の結果を比較した．その後，評判傾向の変化点におけるトピックを抽出（実験 3）し，未知データに対して今回抽出した回帰式をもとに調査対象の増減を推定した（実験 4）．以降，本論文で使用したデータセット，本手法によるコンテンツの切分け，本手法と実際の調査との相関，チャンキングを用いたトピック抽出，未知データに対する推定の順に述べる．

5.1　データセット

本手法の有効性を評価するために，本論文では政治とテレビドラマに関するコンテンツについて本手法を適用した．政治に関しては 2010 年 6 月 24 日から同年 12 月 10 日までを期間とし，同期間における内閣名，政党名を含む twitter のコンテンツ 1,469,594 件を解析対象とした．同様にテレビドラマに関しては 2010 年 11 月 1 日から同年 12 月 23 日までを期間とし，同期間におけるドラマタイトル名を含む twitter のコンテンツ 474,137 件を解析対象とした．

5.2　実験 1：本手法によるコンテンツの切分け

本論文では投稿者の意見を抽出するために特定のキーワードの位置と twitter の機能を用いてユーザコンテンツを切り分けた．機能を用いた投稿の解析対象の位置が適切であったかを検証するために，機能を使用した場合における投稿者の意見の位置の内訳を調査した．今回の調査では政治に関する意見を含み，機能を使用したコンテンツ 1000 件について調査し，投稿者の意見が記されている場所を人手で集計した．その内訳は，機能直前の文字列 323 件，機能直後の文字列 593 件，機能直前・直後以外の文字列 84 件であった．

表 2　解析対象の切り分けの内訳（政治）
Table 2　Result of analysis target distribution. (Politics)

	図 3 中の 1	図 3 中の 2	図 3 中の 3	図 3 中の 4	図 3 中の解析対象外
件数	102224	473340	143877	622703	127450
割合（%）	6.96	32.21	9.79	42.37	8.67

表 3　解析対象の切り分けの内訳（テレビドラマ）
Table 3　Result of analysis target distribution. (TV dramas)

	図 3 中の 1	図 3 中の 2	図 3 中の 3	図 3 中の 4	図 3 中の解析対象外
件数	8557	70984	17752	301590	75254
割合（%）	1.80	14.97	3.74	63.61	15.87

表 4　各手法の重回帰分析結果（政治）
Table 4　Multiple regression results for p/n determination vs. sentiment analysis. (Politics)

	重相関 R	決定係数 R^2	自由度調整済決定係数 R'^2	有意 F
p/n 判定	0.50	0.25	0.22	0.0003
センチメント分析	0.70	0.49	0.38	0.0001
p/n 判定+センチメント分析	0.82	0.68	0.60	5.5E-08

表 5　各手法の重回帰分析結果（テレビドラマ視聴率）
Table 5　Multiple regression results for p/n determination vs. sentiment analysis. (TV dramas)

	重相関 R	決定係数 R^2	自由度調整済決定係数 R'^2	有意 F
p/n 判定	0.46	0.34	0.26	0.001
センチメント分析	0.74	0.58	0.33	0.0009
p/n 判定+センチメント分析	0.80	0.65	0.56	2.91E-06

このように機能を用いているコンテンツを解析する際には，機能の前後に着目することで投稿者の意見を抽出できたことが分かった．また，表 2，表 3 は解析対象の切分けの内訳を示している．表 2，表 3 に示すように，解析対象外となるものは 8.67%（政治），15.87%（テレビドラマ）の割合を占めていた．こららの解析対象外となるコンテンツは，例えばコンテンツを「AAA RT @UserNameA BBB QT @UserNameB CCC」や「RT @UserNameA BBB QT @UserNameB CCC」とした場合，「CCC」にキーワードが存在しているコンテンツを示している．前述した機能を使用した場合における投稿者の意見の内訳に示したように，機能直前・直後以外の文字列である「CCC」には投稿者の意見が含まれていない可能性がある．また，「AAA」や「BBB」を解析した図 3 中の 1〜3 に当たるものは，48.96%（政治），20.51%（テレビドラマ）の割合を占めていた．解析対象によって違いがあるものの，これらの結果により twitter の機能を利用することで，ユーザの意見が反映されている部分を解析対象にすることができたと考えられる．

5.3　実験 2：本手法と実際の調査との相関

また，本手法と実際の調査との相関を評価するために，正解データを政治に関しては 2010 年 6 月 24 日から同年 12 月 10 日までの全国紙の世論調査を使用した．重回帰分析には，式 (1) に示したように目的変数 Y_t に支持率の増減度，説明変数 $X_{[t-1,t]}$ に調査日間における評価表現の出現頻度の平均の増減度を入力とした．同様にテレビドラマに関しては 2010 年 11 月 1 日から同年 12 月 23 日までのビデオリサーチ社が調査したテレビ視聴率を使用した．それぞれに対して p/n 判定，センチメント分析それぞれ単体で重回帰分析をした場合と，両者を併用した場合を比較した．

表 4 は政治，表 5 はドラマに関するコンテンツについて，それぞれ各手法の重回帰分析を実施した結果である．まず，実施した重回帰分析が有意かどうかを判断する指標である有意 F に着目すると，表 4，表 5 に示すようにどの手法も 0.05 以下であることから有意であるという結果が得られた．また，表 4 に示すように，得られる回帰式の当てはまりの度合を示す自由度調整済決定係数 R'^2 に着目すると，政治に関するコンテンツに対しては，p/n 判定 (0.22)，センチメント分

析 (0.38)，p/n 判定とセンチメント分析の併用 (0.60) であった．同様に表 5 に示すようにテレビドラマに関するコンテンツに対しては，p/n 判定 (0.26)，センチメント分析 (0.33)，p/n 判定とセンチメント分析の併用 (0.56) であった．これらの結果から，p/n 判定単体での結果よりも，センチメント分析単体が精度が良く，更に両手法を併用した場合が精度が高いことが確認できた．また，p/n 判定に用いた日本語評価極性辞書とセンチメント分析に用いた感情表現辞書の各辞書に共通する語が占める割合を調査したところ，日本語評価極性辞書 (3.71%)，感情表現辞書 (8.92%) であったことから，両辞書は互いに異なる語を収録した辞書であり，センチメント分析は p/n 判定では抽出することができない評価情報を抽出することができ，その結果，評判傾向抽出の精度が向上したと考えられる．

次に，変数減少法を用いて今回の調査対象である世論調査の増減，及びテレビドラマ視聴率の増減に影響を与える評価表現を特定した．表 6，表 7 は変数減少法の適応後の p/n 判定とセンチメント分析を用いた重回帰分析結果を示している．表 6 に示すように政治に関するコンテンツでは，有意 $F(1.08E^{-09})$，自由度調整済決定係数 $R'^2(0.60)$，表 7 に示すようにテレビドラマに関するコンテンツでは，有意 $F(3.46E^{-07})$，自由度調整済決定係数 $R'^2(0.57)$ の結果を得ることができた．また，各説明変数が有意であるかどうかを判断する指標である p 値に着目すると，表 6，表 7 の各説明変数の p 値は 0.05 以下であることから，これらの評価表現が調査対象である支持率の増減度，視聴率の増減度に影響を与えていることが分かった．

5.4 実験 3：チャンキングを用いたトピック抽出

次に提案手法のチャンキングを適用して，評判傾向の変化点におけるトピックを抽出した．本実験では，期間中における政治とテレビドラマの時間的変化の比較的大きな変化点に対してトピックを抽出した．表 8 は政治の変化点，表 9 はテレビドラマの変化点におけるチャンキングを用いたトピック抽出結果を示したものである．これらのトピックは実験 1 で得られた回帰式から特定した変化点におけるトピックであるた

表 6　支持率の増減に影響を与える評価表現
Table 6　The evaluate expressions affecting to the poll.

定数項 α'		0.749
回帰係数 $\beta' i$	p	$0.077^{\star\star}$
	n	$-0.078^{\star\star}$
	喜	$0.083^{\star\star}$
	怖	$0.051^{\star\star}$
	恥	$-0.053^{\star\star}$
	厭	$-0.062^{\star\star}$
	昂	$-0.010^{\star}$
重相関 R		0.81
決定係数 R^2		0.65
自由度調整済決定係数 R'^2		0.60
有意 F		1.08E-09

$\star\star$:$p < 0.01$，$\star$:$p < 0.05$

表 7　視聴率の増減に影響を与える評価表現
Table 7　The evaluate expressions affecting to the TV dramas.

定数項 α'		0.880
回帰係数 β'_i	p	$0.151^{\star\star}$
	n	$-0.104^{\star\star}$
	哀	$-0.073^{\star\star}$
	怖	$0.0811^{\star\star}$
	恥	$-0.025^{\star}$
	厭	$-0.061^{\star\star}$
	安	$0.019^{\star}$
	驚	$0.040^{\star\star}$
重相関 R		0.79
決定係数 R^2		0.63
自由度調整済決定係数 R'^2		0.57
有意 F		3.46E-07

$\star\star$:$p < 0.01$，$\star$:$p < 0.05$

表 8　変化点におけるチャンキングを用いたトピック抽出結果（政治）
Table 8　Chunking results. (Politics)

変化点	トピック
7 月 10 日 09:00～09:59	参院選の民主敗北必至の情勢
9 月 11 日 13:00～13:59	理解できないよね、民主党ってのは
11 月 5 日 11:00～11:59	【尖閣ビデオ流出問題】「菅政権は末期症状だ」「国が公開すべきだった」

表 9　変化点におけるチャンキングを用いたトピック抽出結果（テレビドラマ）
Table 9　Chunking results. (TV dramas)

変化点	トピック
12 月 8 日 00:00～00:59	『医龍 3』予告篇 OA 後の反響がすごい
12 月 15 日 14:00～14:59	「フリーター、家を買う。」が 18.6%の高視聴率を記録
12 月 21 日 15:00～15:59	フジ系今夜 9 時からは「フリーター、家を買う。」最終回！

め，評判傾向の変化に影響を与えた可能性がある．特に表 8 の 11 月 5 日の抽出結果，表 9 の 12 月 8 日，12 月 21 日のトピックに関しては，政治に関しては民主党の支持率が減少し，テレビドラマに関しては各ドラマの視聴率が上昇した原因であると考えられる．しかしながら，その一方で表 8 の 9 月 11 日の抽出結果のように投稿者の意見が抽出される場合もあった．これは，本手法では出現回数の多い形態素群をトピックとしているため，意見とトピックを判別することができなかったためだと考えられる．

5.5 実験 4：未知データに対する推定

実験 2 で抽出した支持率に関する回帰式を用いて未知データに対して支持率の増減を推定した．今回使用した未知データは，2010 年 12 月 11 日から 2011 年 1 月 31 日までのコンテンツ 335,673 件を利用し，正解データとして同期間における全国紙の世論調査結果を利用した．今回の実験では回帰式に世論調査の各区間における評価表現の増減値を入力し，支持率の増減（前回の世論調査から上昇か下降か）を推定した．この実験では，内閣及び 7 つの政党の支持率の増減（上昇か下降か）の推定を計 80 回試行し，約 66%の確率（全 80 回中 53 回正解）で一致した．これにより，評判傾向予測システムの構築の実現に向け本手法が未知データに対する推定に関しても一定の結果を得ることが可能であると考えられる．しかしながら，より精度の高い推定結果を得るために評価表現の抽出法の改善などが必要であると考える．

6. 関 連 研 究

本章では，評判傾向の時間的変化を抽出するサービス，センチメント抽出技術，マイクロブログを用いた研究について紹介する．

まず，評判傾向の時間的変化を抽出するサービスとして kizasi [18] がある．kizasi ではクローラと呼ばれる Web ページを自動取得するエージェントを利用して大量のブログ記事を逐次取得し，そのブログ記事中に表れる単語の出現頻度とその推移から「kizasi 度」と呼ばれる単語の注目度を算出する．BLOGRANGER [19] は，仮想の地図上にインターネット上の複数のブログで話題になっているキーワードを配置し，位置関係，高さ表示によりその話題の関連性，活発さを表現してユーザに対して情報を提供するサービスである．これらのサービスでは，特定のキーワードと共起する語の変遷や関連性により評判傾向の時間的変化を抽出しているがユーザコンテンツのセンチメントには着目していない．

センチメントに着目した研究としては，情報発信者（話し手，書き手）の感情を推定しようという研究 [20], [21] が盛んに行われている．ただし，これらの研究は情報発信者の感情に応じた応答を生成あるいは支援することにより，システム（例えばロボットやアバタなど）とのコミュニケーションや通信システムを介した，人どうしのコミュニケーションをより円滑にしようというものであり，コンテンツの感情表現を抽出しようというものでない．一方，コンテンツの感情表現を抽出する研究では，複数ニュースサイトの差異情報可視化の研究がある [22]．この研究では，ニュースサイトごとの観点の相違を抽出しており，ユーザコンテンツに着目していない点が本論文と異なる．

また，マイクロブログを用いた研究は，マイクロブログにおけるの有力ユーザを抽出する研究 [23]，有用性のあるコンテンツを推薦する研究 [24] などがある．これらの研究では，マイクロブログの特徴であるユーザ間の「弱い」関係とユーザ間のコミュニケーションに着目している．またマイクロブログのリアルタイム性に着目した研究では，政治家のテレビ討論中におけるユーザの議論の動向を抽出する研究 [26] や，特定の場所に関してリアルタイムな情報を提示する研究などがある [28]．本論文のようにマイクロブログから特定の事象を抽出する研究では，各ユーザをセンサに見立てて，ユーザコンテンツから地震の震源地や台風の予想進路を抽出する研究 [25] や，災害時における情報の広がり方を調査した研究 [27] などが挙げられる．また政治に関するコンテンツを対象とした研究としては，議会選挙時における政党別の支持率を政党名を含むコンテンツ数のみで算出し，実際の選挙結果とを比較する研究がある [29]．本論文では，マイクロブログの構造と特性を用いた評価情報の抽出法を提案し，重回帰分析によって調査対象のモデル化と影響を与える評価表現を特定している点が異なる．

7. む す び

本論文では，評判傾向の時間的変化と，その原因を抽出する評判傾向抽出エージェントを実現するために，マイクロブログのコンテンツに p/n 判定とセンチメント分析を実施し，重回帰分析を用いて評判傾向の変化点を求め，変化点におけるトピックを抽出する手法を提案した．また，政治及びテレビドラマに関するコン

テンツを対象に本手法を適用し，政治（自由度調整済決定係数 R'^2（p/n 判定単体：0.22，提案手法：0.60)），テレビドラマ（自由度調整済決定係数 R'^2（p/n 判定単体：0.26，提案手法：0.56)）の結果を得ることができた．これにより従来の評価判定法である p/n 判定にセンチメント分析を組み合わせることで p/n 判定単体よりも人手による調査と相関の高い時系列変化を抽出できることが確認できた．

今後の課題として，抽出したトピックの判別法がある．本論文では，出現回数が多い順にトピック候補をランキングしトピックとして決定した．今後は，これらを意見とトピックとに判別し，抽出したトピック後の調査の影響を検証することで，変化の原因抽出の精度向上を目指したい．また，本論文では政治，及びテレビドラマに関するコンテンツに本手法を適用し一定の結果を得たが，今後は得られた回帰式を元により高い精度で調査対象（支持率，テレビ支持率）を推定する手法を検討していきたいと考えている．

謝辞　本研究を遂行するにあたり，研究の機会と議論・研鑽の場を提供して頂き，御指導頂いた国立情報学研究所/東京大学本位田真一教授をはじめ，活発な議論と貴重な御意見を頂いた研究グループの皆様に感謝致します．

文　献

[1] 松本正生，“特集世論調査方法の再検証—「総合調査学」へ向けて，” 行動計量学，vol.35, no.1, pp.1–3, 2008.
[2] 総務省，“ブログ・SNS の現状分析及び将来予測，” 総務省報道資料，2005.
[3] M. Ebner and M. Schiefner, “Microblogging - More than fun?,” Proc. IADIS Mobile Learning Conference, Inmaculada Arnedillo Sanchez and Pedro Isaias ed., pp.155–159, Algarve, Portugal, 2008.
[4] 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, pp.201–241, 2006.
[5] K. Dave, S. Lawrence, and D.M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” International World Wide Web Conference, Proc. 12th International Conference on World Wide Web, pp.519–528, 2003.
[6] 飯田　龍，小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一，“意見抽出を目的とした機械学習による属性-評価値対同定，” 情報処理学会自然言語処理研究会予稿集，NL-165-4, pp.21–28, 2005.
[7] 福原知宏，宇津呂武仁，中川裕志，武田英明，“複数の言語で記述されたブログ記事を対象とした言語横断型関心解析システム，” The 21st Annual Conference of the Japanese Society for Artificial Intelligence, CD-ROM (2F4-3), 2007.
[8] 総務省情報通信政策研究所（IICP）調査研究部，“ブログの実態に関する研究の結果，” 2008.
[9] twitter, http://twitter.com/
[10] A. Java and X. Song, “Why we Twitter: Understanding microblogging usage and communities,” WebKDD and 1st SNA-KDD Workshop, 2007.
[11] 枝　洋樹，林　信行，小林弘人，津田大介，武田　徹，高須賀宣，岡野原大輔，片瀬京子，高橋秀和，亀津　敦，Twitter の衝撃，日経 BP 社出版局，2009.
[12] 加藤恒明，松下光範，時系列情報の抽出と可視化に基づく情報アクセスためのマルチモーダルインタフェース—情報編纂の基礎技術に向けて，人工知能学会，2007.
[13] 中村　明，感情表現辞典，東京堂出版，1993.
[14] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一，“意見抽出のための評価表現の収集，” 自然言語処理，vol.12, no.3, pp.203–222, 2005.
[15] CaboCha, http://chasen.org/taku/software/cabocha/
[16] MeCab, http://mecab.sourceforge.net/
[17] 高木俊宏，注連隆夫，土屋雅稔，“機械学習を用いた日本語機能表現のチャンキング，” 言語処理学会，J. Natural Language Processing, vol.14, no.1, pp.111–138, 2007.
[18] kizasi, http://kizasi.jp/
[19] K. Fujimura, H. Toda, T. Inoue, N. Hiroshima, R. Kataoka, and M. Sugizaki, “BLOG RANGER-A multi-faceted blog search engine,” International World Wide Web Conference, Proc. 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006.
[20] 倉石英俊，柴田義孝，“個人モデルを用いた表情分析/合成による感情通信システム，” 情処学研報，マルチメディア通信と分散処理，74-14, 1996.
[21] 松本和幸，黒岩眞吾，任　福継，“感情計測システムについて，” 信学技報，NLC2003-10, 2003.
[22] 濱砂佳貴，河合由起子，熊本忠彦，田中克己，“センチメントマップによる複数ニュースサイトの差異情報可視化手法の提案，” 第 19 回データ工学ワークショップ (DEWS2008) 論文集，B6-4, 2008.
[23] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twitt er-Rank: Finding topic-sensitive influential twitterers,” Web Search and Web Data Mining, Proc. Third ACM International Conference on Web Search and Data Mining, pp.261–270, 2010.
[24] 岩木祐輔，アダム ヤトフト，田中克己，“マイクロブログにおける有用な記事の発見支援，” 電子情報通信学会・日本データベース学会・情報処理学会第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009), A6-6, 2009.
[25] T. Sakaki, M. Okazaki, and Y. Matsuo, “Ear thquake shakes twitter users: Real-time event detection by social sensors,” International World Wide Web Conference, Proc. 19th International Conference on World Wide Web, pp.851–860 , 2010.
[26] D.A. Shamma, L. Kennedy, and E.F. Churchill,

"Tweet the debates: Understanding community annotation of uncollected sources," International Multimedia Conference, Proc. First SIGMM Workshop on Social Media, pp.3–10, 2009.

[27] S. Vieweg, A.L. Hughes, K. Starbird, and L. a Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," Conference on Human Factors in Computing Systems, Proc. 28th International Conference on Human Factors in Computing Systems, pp.1079–1088, 2010.

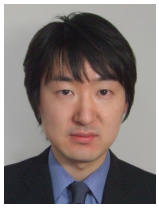
[28] 松村飛志，安村通晃，"街に着目した Twitter メッセージの自動収集と分析システムの提案と試作，" 情報処理学会インタラクション，2010.

[29] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," International AAAI Conference, Proc. Fourth International AAAI Conference on Weblogs and Social Media, 2010.

（平成 23 年 1 月 13 日受付，5 月 23 日再受付）

橋本　和幸

2009 電通大・電気通信・情報通信卒．現在，同大大学院情報システム学研究科修士課程社会知能情報学専攻在学．



中川　博之　（正員）

1997 阪大・基礎工・情報工学卒．同年鹿島建設（株）に入社．2007 東京大学大学院情報理工学系研究科修士課程了，2008 同大学院博士課程中退．同年より電気通信大学助教，現在に至る．エージェント及び自己適応システム開発手法の研究に従事．情報処理学会，IEEE CS 各会員．

田原　康之

1991 東京大学大学院理学系研究科数学専攻修士課程了．同年（株）東芝入社．1993〜1996 情報処理振興事業協会に出向．1996〜1997 英国 City 大学客員研究員．1997〜1998 英国 Imperial College 客員研究員．2003 国立情報学研究所入所．2008 より電気通信大学准教授．博士（情報科学）（早稲田大学）．エージェント技術，及びソフトウェア工学などの研究に従事．情報処理学会，日本ソフトウェア科学会会員．

大須賀昭彦　（正員）

1981 上智大・理工・数学卒．同年（株）東芝入社．同社研究開発センター，ソフトウェア技術センターなどに所属．1985〜1989（財）新世代コンピュータ技術開発機構（ICOT）出向．2007 より，電気通信大学大学院情報システム学研究科教授．工博（早稲田大学）．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞受賞．情報処理学会，日本ソフトウェア科学会，人工知能学会，IEEE CS 各会員．