社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143 (34)
2007／7／3

ブログ記事からのトピック別評判情報変遷パタンの抽出手法について

戸田 智子† 鎌田 基之† 黒田 晋矢† 福田 直樹†† 石川 博††

† 静岡大学大学院情報学研究科 〒432–8011 静岡県浜松市中区城北3–5–1
†† 静岡大学情報学部情報科学科 〒432–8011 静岡県浜松市中区城北3–5–1
E-mail: †{gs07037,gs07016,gs07023}@s.inf.shizuoka.ac.jp, ††{fukuta,ishikawa}@inf.shizuoka.ac.jp

あらまし 現在，ブログの爆発的普及により，多くのユーザがブログを活用している．ブログに記述される個人の意見・感想などの評判情報が注目を集めている．本論文では，特に，評判情報を抽出する対象として，製品等ではなく，ブログ記事から抽出されたニュースやイベント等のトピックに焦点をあてる．トピックに対して評判情報を抽出することにより，実世界で起きたイベントに対する社会の動向を伺うことが可能であると考えられる．本論文では，ブログ記事群からトピックを抽出し，抽出したトピックに対して評判情報の変遷を可視化する手法を提案する．
キーワード ブログ文書，クラスタリング，変遷パタン，評判情報

On Extraction of Reputation Information Patterns of Each Topic from Blog Articles

Tomoko TODA†, Motoyuki KAMADA†, Shinya KURODA†, Naoki FUKUTA††, and Hiroshi ISHIKAWA††

† Graduate School of Informatics, Shizuoka University Johoku 3–5–1, Nakaku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
†† Department of Computer Science, Faculty of Informatics, Shizuoka University Johoku 3–5–1, Nakaku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
E-mail: †{gs07037,gs07016,gs07023}@s.inf.shizuoka.ac.jp, ††{fukuta,ishikawa}@inf.shizuoka.ac.jp

Abstract Today, with explosive spread of blogs, people have their own blogs and use them frequently. Reputation information described in blogs as personal opinions often has a significant impact on other people in purchasing goods and services. We focused on general topics from extracted from blog articles as reputation information. Analysing such extracted reputation information will allow us to capture various changes and transitions of responses in the society for the events occurred in the real world. In this paper, we propose a method that automatically extracts and visualizes transition patterns of reputation information on topics from a set of blog articles.
Key words Blog,Clustering,Transition Patterns,Reputation Information

1. はじめに

現在，新たな情報発信の手段としてブログが注目されている．ブログの普及により，多くのユーザが個人の意見を発信できるようになり，世の中のさまざまな人の意見が記述されている．総務省の調査[1]によると，2005年末ではブログ利用者は約335万人，ブログ閲覧者は約1,651万人いるとされ，2007年末には利用者が約782万人，閲覧者が約3,455万人にまで達すると予測されている．

現在のブログの特徴には，ウェブ上における個人の日記という側面と，特定のニュースやイベントに対する個人の意見を表現するメディアの1つという側面がある．特定のニュースやイベントに対する個人の意見を表現するメディアであるという側面においては，特定の対象に関する評判情報の扱いが重要視されてきている．

評判情報を抽出する研究においては，何らかの製品や商品に焦点を当てた研究が多く存在している．しかし，製品に関する評判情報のみでなく，製品や商品に直結していないようなもの(イベントやサービスなど)に対して，どのような反応があったのかを知りたい場合もある．

本論文では，こういった，あるニュースやイベントに対する反応・話題を**トピック**と呼ぶこととする．こういったトピックを対象とする際には，その評判情報がどのように変化してきたかということもあわせて重要となってくると考えられる．

我々は,[2] において，ブログ記事群よりトピックを抽出し，抽出したトピックに対して，投稿されたタイムスタンプと記事数に基づいて，トピックの変遷を抽出し，可視化する手法を提案した．本論文では,[2] の手法を用いて抽出したトピックに対して，評判情報の変遷を抽出していく手法を提案する．

本論文では,あるトピックに対する評判情報の変遷を抽出し，可視化する手法を提案する．ブログ記事から抽出したトピックに対して，その肯定・否定の時系列上の変化を可視化することにより，製品や商品に直結していないようなものに対する評判情報を抽出することを可能とし，その動向も容易に抽出可能とすることを目指す．

2. 関連研究

ブログの個人性や時系列性に着目し，ブログ内での評判情報の抽出や話題の変遷の抽出を行う研究が，これまでに提案されてきている．

ブログの話題抽出や話題変遷の抽出技術を応用したサービスとして，多くのブログ著者の間で記述されている話題について関連語などを表示し，それに併せてその話題が投稿された記事数を時間軸を用いて可視化しているサービスが提供されている [3]．これは，着目したトピックに関して記述された記事数のみの変遷を可視化しているものであり，本研究ではトピック中の評判情報の変遷を可視化することを目的としている点において，このサービスとは異なっている．

トピック抽出手法としては，burst の検出によるものが挙げられる．手法 [4] 及び [5] では，ある語に対し，その語が出現する時間間隔の定常状態を求めておき，その時間間隔よりも短い間隔で語が出現しているとき，その語をトピックに関連する語として抽出する．手法 [4] 及び [5] では，急激に話題になったようなトピックの抽出を目的としたものであり，ほとんど変化なく取り扱われているトピックについてはうまく抽出できない．本研究では，急激に話題になったようなトピックだけではなく，ほとんど変化なく扱われているようなトピックに対しても評判情報を抽出することを目標としているため，手法 [4] 及び [5] では目的が達成されない．

Web ページやブログを対象とした評判情報の抽出として立石らの手法 [6] が挙げられる．立石らは，インターネットからの評判情報検索として，商品名からその商品に関する評判情報を抽出することを行っている．この手法では，商品の種類によって評価を示す語が異なることから，あらかじめ商品をカテゴリ化しておき，そのカテゴリごとに評価表現辞書を作成しておく．入力された商品名が出現した箇所の近くに該当する商品カテゴリの評価表現辞書に登録されている語があるかどうかで評判情報を判定し，抽出している．

藤村らは，掲示板から評価表現を抽出し，肯定的な表現と否定的な表現において，それらの差分をとることにより，肯定的な表現と否定的な表現を抽出する手法 [7], [8] を提案している．

また，ブログからの評判情報抽出を行った研究としては，以下のようなものがが挙げられる．鈴木らは，評価表現を対象・評価部分・評価の三つ組みとして抽出し，この評価表現を用いて肯定・否定または非評価を判断する手法 [9] を提案している．

霜田らは，ブログから評判情報を抽出するための辞書をユーザに作成してもらい，興味対象を含んでいるキーワードと評価表現を用いて，評判情報が記述されているブログ記事をユーザに提供する手法 [10] を提案している．この手法では，ユーザに評価語とその肯定・否定を入力させることにより，よりユーザの意図にあった評判情報を抽出している．

本論文で提案する手法は,ブログ記事において記述されているトピックを評判情報抽出の対象としている点，また，複数のトピックの評判情報の変遷を同時に可視化することにより，関連するトピック間において評判情報の変遷の相関などを抽出可能な手法である点で，これらの関連研究と大きく異なる．

3. トピックと評判情報の抽出

本手法は，トピック抽出フェーズと評判情報抽出フェーズの 2 つに，大きく分けられる．本手法では，トピック抽出フェーズで得られた結果に対し，評判情報抽出フェーズ，及び，評判変遷抽出フェーズによってトピックごとの評判情報の変遷の抽出を行う．

トピック抽出フェーズでは，ブログ記事群に対してクラスタリングを行うことにより，ブログ記事の集合としてトピックを抽出する．次に，評判情報抽出フェーズでは，各トピックに含まれる記事から，評価表現を抽出する．抽出された表現の肯定・否定に基づいて，ブログ記事を肯定・否定に分類することにより，ブログ記事から評判情報を抽出する．抽出した評判情報とブログ記事が投稿されたタイムスタンプに基づいて，評判情報の変遷を可視化する．

3.1 トピックの抽出

ブログ記事に形態素解析を行い，名詞・動詞・形容詞を抽出する．抽出した名詞により，文書ベクトルを生成する．生成した文書ベクトルに基づいて，ブログ記事のクラスタリングを行う．得られたクラスタをトピックとし，トピック中から話題として扱えそうなトピックのみをクラスタに含まれる記事数によって，自動的に選択することによって，ブログ記事群からトピックを抽出する．

ブログ記事の，それぞれタイトルと本文に対して形態素解析を行う．形態素解析ツールとしては，Sen [11] を用いる．Sen によって形態素解析を行った結果から,名詞・動詞・形容詞を抽出する．抽出した語のうち，名詞のみを用いて文書ベクトルを生成する．ここで，文書ベクトル作成に用いる語には，非自立語，接尾語，数，代名詞を除くものとする．

連続して出現している名詞はもともと複合語である名詞が分割されたものと考えられる．これらを結合し 1 つの名詞とみなしたほうが，結果が良好となる場合が考えられる．ブログ中においては，口語体のような記述がなされている記事や，句読点があまりつけられていないような記事も多く存在するため，複合語以外にも名詞が連続して出現する場合がある．名詞が連続する全ての場合に結合を行うと，記事中に現れる特徴的な複合語以外にも，不必要に多くの語が登録されることとなってしまう．本研究では，名詞の結合を行う場合は，連続して現れる名

表 1　名詞の結合有，結合無の例

結合有	空気 ／ 清浄 ／機 (一般＋サ変接続＋接尾)
	健康 ／ 食品 (形容動詞語幹＋一般)
	小泉 ／ 純一郎 (人名姓＋人名名)
結合無	明日 ／ テスト (副詞可能＋一般)
	3 ／ 回 ／ 目 (数＋接尾＋接尾)

詞が，副詞可能，非自立語，数，代名詞以外の場合のみに行うこととする．結合を行う語のうち，接尾語に関しては，直前に出現している名詞に結合する場合のみ登録することとし，単独での登録は行わないようにする．Sen では人名は**姓名詞**と**名名詞**に分割されるが，**姓名詞**と**名名詞**が連続して出現するような場合には，一人の人物の姓および名であると考えられるため，姓と名を結合した**人名名詞**も，**姓名詞・名名詞**とあわせて，文書ベクトル中に登録することとする．名詞が連続して出現した際における，結合を行う場合と行わない場合の例を表 1 に挙げる．

ブログ記事ごとの文書ベクトルの各要素に対しては一般的な TFIDF を用いた重み付けを行う．あるブログ記事 E における語句 t の重み w_E^t は式 (1) によって求める．

$$w_E^t = \frac{log(tf(t,E)+1))}{log(M)} * log(\frac{N}{df(t)}) \quad (1)$$

ここで，$tf(t,E)$ はブログ記事 E 中に語句 t が出現する頻度，$df(t)$ は全ブログ記事中において単語 t が出現しているブログ記事数，N は形態素解析を行ったブログ記事の総数，M はブログ記事 E より抽出された単語の種類数を示す．

作成した文書ベクトル群に対して，凝集型の階層的クラスタリングを行う [12]．凝集型の階層的クラスタリングでは，初期段階としてそれぞれの要素を 1 つのクラスタとしてみなし，それらを併合していくことによってクラスタを生成していく．最終的にはクラスタ数が 1 になるまで併合されていくが，クラスタリング終了の閾値を設けることにより，任意の大きさのクラスタを生成可能である．ここでは，同じような内容を記述している記事をまとめることにより，トピックとそのトピックに関連する語を抽出することを目的としている．したがって，生成されるクラスタがあまり大きくなりすぎないように閾値を決定する必要がある．また，計算量の軽減のため，クラスタリングに用いる類似度の算出にはベクトル中の全ての語を使用するのではなく，その語の TFIDF 値がある閾値以上のもののみを使用することとする．

文書ベクトルの一般的な類似度算出式では，それぞれのベクトルの大きさによる正規化を行っている (式 2)．ブログの特徴のひとつとして，ひとつの記事中に二つ以上のトピックについて言及しているような場合が多く存在することが挙げられる．これは，長い記事であってもそのトピックに関することが必ずしも多く記述されているわけではないということを意味する．

今回は，ブログ記事間の類似度を算出することにより，同一のトピックについて記述している記事を検出することを目的としている．1 つの記事で，複数のトピックについて記述されているような場合においても，記事の長さに影響されず抽出が行えるほうが本研究では望ましい．したがって，ブログを対象とする場合には，ベクトルの大きさによって正規化しない式 (式 3) のほうが良い結果が得られる可能性がある．

$$sim(E_i,E_j) = \frac{w_i^1 w_j^1 + \ldots + w_i^m w_j^m}{\sqrt{(w_i^1)^2 + \ldots + (w_(i)^m)^2} * \sqrt{(w_j^1)^2 + \ldots + (wj^m)^2}} \quad (2)$$

$$sim(E_i,E_j) = w_i^1 w_j^1 + \ldots + w_i^m w_j^m \quad (3)$$

予備実験として，ブログ記事 E_i，E_j における類似度を式 2 と 3 において比較を行った．呼び実験の結果，式 3 によって求めた類似度が式 2 に比べ，トピックの抽出という目的に対してより精度がよいことが確認された．よって，本研究では，ブログ記事 E_i, E_j に対する類似度は式 3 で求めることとする．

また，クラスタ間の類似度を求める際には，最長距離法を用いて行う．最長距離法とは，クラスタ間の類似度を算出する際に，そのクラスタに属する組のうち，最も類似度の低い組み合わせの類似度をそのクラスタ間の類似度とする手法である．この手法では，比較的細やかなクラスタが生成されると考えられるため，トピック抽出に関しても話題の混濁が起こりにくいと考えられる．よって，本研究では，最長距離法を用いてクラスタ間の類似度を算出することとする．

ブログ記事に対するクラスタリングにより生成された各クラスタのうち話題として扱えそうなクラスタを選択する．話題として扱うためには，ある程度の記事間で共有されているトピックでなくてはならない．1 つのトピックのクラスタに含まれる記事がごく少数 (特に 1 つしか含まれない場合) であるようなトピックは，話題としては適さないと考えられる．この基準により，話題として適していると判断されたクラスタのみを対象として，評判情報を抽出することとする．

3.2　評判情報の抽出

トピック抽出フェーズにおいて抽出されたトピックよりそのトピックの評判情報を抽出する．トピック中に含まれる各記事に対して，その記事に記述されている内容の肯定・否定を分類する．

まず，トピック中に含まれる各記事に対して，評価表現が含まれているかどうかを調べる．評価表現が含まれていれば，その評価表現の肯定・否定により記事の肯定・否定を分類する．1 つの記事の内容の肯定・否定に関するスコアは，肯定評価表現，否定評価表現，それぞれの出現回数に基づいて算出する．

3.2.1　評価表現の抽出

ブログ記事から評価表現を抽出する．評価表現はトピックに

表 2　評価表現語句例

肯定表現語句	良い
	いい
	安い
否定表現語句	悪い
	わるい
	高い

表 3　評価反転例

評価反転有	良くない
	悪くはない
評価反転無	良くない訳ではない
	悪くないこともない

よって異なると考えられる．したがって，目的とするトピックに応じて評価辞書を設定する必要がある．

本論文では，評価表現はユーザから与えられた語とする．ユーザが任意に設定した形容詞により，ユーザからの評価辞書を構築する．この評価辞書に含まれるものとマッチしたものを評価表現として抽出する．本論文では，ユーザから与えられた評価表現を評価表現語句と呼ぶこととする．評価表現語句の例を表 2 に示す．

抽出された評価表現語句が評価反転表現を伴って使用されている場合は，その評価表現語句の評価を反転させる必要がある．ここで言う評価反転表現とは，「不〜」や「〜ない」といった語のことを示す．こういった評価反転表現はそれぞれ，「不〜」は名詞の接頭語として，「〜ない」は形容詞や動詞の打ち消しとして出現する．本論文では，評価表現語句として形容詞を想定しているために，評価表現語句を反転させる否定表現としては，「〜ない」のみを対象とすることにする．

文章中での「〜ない」の出現パタンとしては，形容詞＋「ない」，動詞＋「ない」，(形容詞または動詞) ＋「ない」＋非自立語名詞＋「ない」が挙げられる．このうち，形容詞＋「ない」，動詞＋「ない」というパタンで出現する場合は，その直前に出現している形容詞や動詞の単純な否定となっている．そのため，この場合にはその直前に出現している形容詞や動詞の評価を反転させることとする．一方，(形容詞または動詞) ＋「ない」＋非自立語名詞＋「ない」のパタンである場合は，その直前に出現している形容詞や動詞に対して二重否定となる．したがって，この場合には，その直前に出現している形容詞や動詞の評価は反転させないこととする．

評価の反転有り，無し，それぞれの場合にの例を表 3 に示す．

3.2.2　記事の肯定否定の決定

その記事に含まれる評価表現語句によってその記事に記述されている内容の肯定・否定のスコアを算出する．各記事中で肯定表現語句・否定表現語句，それぞれの出現回数に基づいて，各記事のに対し，肯定・否定のスコアを算出する．

あるブログ記事 E において，肯定･否定を示すスコア $score(E)$ は式 4 によって算出する．

$$score(E) = \frac{p\_count * 1 + n\_count * (-1)}{total} \quad (4)$$

ここで，p_count は，ブログ記事 E 中に肯定的な表現が出現した回数，n_coun は，ブログ記事 E 中に否定的な表現が出現した回数を示す．$total$ は，ブログ記事 E 中に評価表現が出現した回数を示し，p_count と n_count の和を示す．記事中に評価表現が存在しない場合は，p_count，n_count ともに 0 であるので，スコアが 0 となる．また，肯定表現，否定表現それぞれの個数が等しい場合も，スコアは 0 となる．これは，肯定・否定それぞれの評価が拮抗している場合も，全体としては未評価に等しいといえると考えられるという仮定に基づいている．よって，本研究では，未評価の場合と評価が拮抗している場合を同等に扱うこととする．

3.3　評判情報変遷の抽出

トピック中の各記事の肯定・否定のスコアと投稿されたタイムスタンプに基づいて，評判情報の変遷を抽出する．

トピック中に含まれる記事のタイムスタンプを抽出する．抽出するタイムスタンプは年月日および時刻とする．トピック中の記事を抽出したタイムスタンプに基づいて分類する．1 日，1 週間，など時間を任意の単位に区切り，その単位ごとにトピック中の記事数とその記事の肯定・否定のスコアの平均を算出する．区切った時間の単位と，その単位ごとの肯定・否定のスコアをグラフ化することにより，各トピックの評判情報の変遷を可視化していく．可視化することによって，トピックの関連する複数のトピックの変遷パタンの比較や，実際に起きたイベントとの比較を行うことができると考えられる．

4.　実　　験

本実験では，データセットとして，クローラで収集した livedoor ブログの記事 28,179 件 (2006 年 7 月 3 日〜2007 年 1 月 31 日) を使用する．これらのうち，「携帯」の語を含む記事 8,787 件に対して，トピックの抽出及び評判情報の抽出，変遷の可視化を行う．

4.1　パラメータの設定

実験に際して，使用する各種パラメータの設定を行う．本実験において，使用するパラメータを次のように設定する．

トピックの抽出において，類似度算出に使用する語句の閾値としては，0.8，凝集型階層的クラスタリングの終了閾値は，3.2 とした．

トピックの選択においては，本実験ではトピックの評判情報の変遷の抽出を目的としているため，トピック抽出の際の，クラスタ内に含む最低の記事数は 10 エントリとした．

トピックの評判情報の変遷の可視化において，変遷を抽出する粒度としては，今回は対象としている期間が短いことから，1 日を基準とすることとした．

4.2　トピック抽出

「携帯」の語を含む記事，8,787 件に対して，トピックの抽出，評価表現の抽出を行った結果，抽出されたトピック数は 63 個であった．また，抽出された各トピック中に含まれる，平均の記事数は 18.7 であった．

抽出されたトピックのうち，スパムや広告の類を除去して得られたもののうち，番号ポータビリティに関するトピックについてを，表4に示す．表中における特徴語とは，それぞれのトピック中において，その語の重みが高い語，上位5語のことを示している．語の重みとは，それぞれ3.1式1にて算出した値の平均のことである．各トピックのラベルとは，クラスタ内に含まれる記事の内容に基づいて，人手によってつけたものである．

表4によると，抽出されたトピックとしては，各携帯電話会社のサービスに関連したトピックや，番号ポータビリティ(MNP)などのトピックなどが抽出されていることがわかる．これは，この期間中に新たな料金プランの発表や，番号ポータビリティなどのイベントが起きたことによると考えられる．

また,「バトン」と呼ばれるものに関するトピックも出現している．したがって，比較的まとまりのあるトピックが抽出できたということが出来ると考えられる．

4.3 評判情報変遷の抽出

表4で得られたトピックのうち，トピックCに対して，評判情報の変遷を可視化したものを図1に，トピックGに対して可視化したものを，図2示す．変遷を抽出した期間としては，2006年10月18日～2006年11月2日の間とした．

図1の結果によると，全体的に比較的肯定的な評判であることがわかるが，10月30日の近辺で急激に否定的な評判に移行していることがわかる．これは，10月28日から3日間にかけて，ソフトバンクの番号ポータビリティの受付が停止したことによるものと考えられる．

図2の結果では，図1とは対照的に，10月30日の近辺でも肯定的な評判となっていることがわかる．トピックGに含まれる記事において，10月30日に投稿されている記事を調べると，この記事から抽出されている評価表現としては，疑問文として記述されているものであった．これは，肯定的な評価表現を用いた疑問文とすることにより，実際には否定的な意味を示唆しているものであり，正しく評価が抽出できていないことがわかる．よって，疑問文中で評価表現語句が出現した場合には，評価表現として扱わない，または，その語の評価表現としての重みを低くし，算出しているスコアに大きな影響を与えないようにする必要があると考えられる．

また，抽出されたトピック全体を通じて，肯定的な評価表現のほうが多くていたことがわかった．今回は評判情報を抽出する際の評価表現に対して，直接的な語を用い，その語の出現回数のみで肯定・否定のスコアを算出している．したがって，抽出された評価表現語句に偏りがある可能性がある．今後，評価表現語句に重みを与えて算出する方法や，異なる評価表現抽出方法にて行った場合も同様の傾向がでるのかについても，検討していきたい．また，異なる分野のトピックに対して行った場合にはどのような傾向がでるのかなどと，あわせて比較・検討していきたい．

4.4 考　　察

抽出されたトピック関しては，同じイベントに対するトピックであっても，異なるトピックとして認識されるものが出現していることがわかる．これは，現段階ではブログ記事1件を1つの文書とみなし，それらに対してクラスタリングを行うことにより，トピック抽出を行っていることによると考えられる．しかし，ブログ記事では1件の記事中に，複数のトピックについて記述されている場合が多く存在する．このことにより，トピックの細分化が起こっていると考えられる．こういった場合，1件の記事を1つの文書としてみなすより，1件の記事を段落ごとに区切り，1つの段落を1つの文書とみなしたほうが良い精度になる可能性があると考えられる．

抽出されたトピック全体を通じて，肯定的な評価表現のほうが多く出現していたことがわかった．今回は評判情報を抽出する際の評価表現に対して，直接的な語を用い，その語の出現回数のみで肯定・否定のスコアを算出している．したがって，抽出された評価表現語句に偏りがある可能性がある．今後，評価表現語句に重みを与えて算出する方法や，異なる評価表現抽出方法にて行った場合も同様の傾向がでるのかについても，検討していきたい．また，異なる分野のトピックに対して行った場合にはどのような傾向がでるのかなどと，あわせて比較・検討していきたい．

評判情報抽出に関しては，評価表現語句の出現回数に基づいてスコアの算出を行っているが，この手法では，1つの評価表現語句に対して，対象が複数並列に存在している場合にはうまく抽出することができない．したがって，単純に評価表現語句の出現数のみに基づくのではなく，それらの評価表現語句が評価している対象の数についても考慮していく必要がある．また，疑問文においては，肯定的な表現語句を用いていても，必ずしも肯定的に捉えているとは限らないことがわかった．よって，評価表現語句が含まれている文の特性 (疑問文，肯定文，否定文など) により，評価表現語句の扱いに対して考慮する必要があると考えられる．

5. おわりに

本論文では，ブログ記事からトピックを抽出し，それらのトピックに対して評判情報の変遷を可視化する手法を提案し，予備的実験を行った．今後の課題としては，大量データへの適用を行い，さまざまな分野のトピックに対しても，評判情報の変遷を抽出することが挙げられる．加えて，長期間のデータへの適用により，変遷を扱う粒度を大きくした際についても検討していく必要がある．関連する複数のトピック間において，評判情報の変遷を比較することにより，関連するトピック間の評判情報の相関などについても検討していく必要がある．

トピック抽出方法に関しては，ブログ記事1件を1つの文書とみなすのではなく，ブログ記事を段落などによって分割し，分割したものを1つの文書とみなしてトピック抽出を行う手法についても検討していく必要がある．

評判情報抽出に関しては，まず，係り受け解析などを用い，1つの評価表現語句が複数を評価の対象としている場合を考慮することが挙げられる．加えて，疑問文や肯定文，否定文などといった文章の特性を考慮していく必要が挙げられる．

謝辞　本研究の一部は科学研究費補助金基盤研究（B)（課題

表 4　抽出されたトピック

トピック	記事数	ラベル	特徴語
トピック A	21	ドコモのサービスについて	ドコモ，携帯電話，パソコン向け，パケット，電源
トピック B	15	料金プランについて	ソフトバンク，通話，料金，定額，プラン
トピック C	14	携帯機種変更について	機種変更，機種，変更，ソフトバンク，灰
トピック D	14	ソフトバンクの CM について	ソフトバンク，ブラッド，ピッド，ディアス，モバイル業界
トピック E	12	MNP au 一人勝ち	DDI，ソフトバンク，NTT ドコモ，川井，執行役員
トピック F	12	ソフトバンク	ソフトバンク，孫，通話，値下げ，ホークス役員
トピック G	11	ソフトバンクの料金プランについて	ソフトバンク，料金，ドコモ，料金プラン，番号役員
トピック H	37	「バトン」について	恋人，バトン，浮気願望，告白，願望

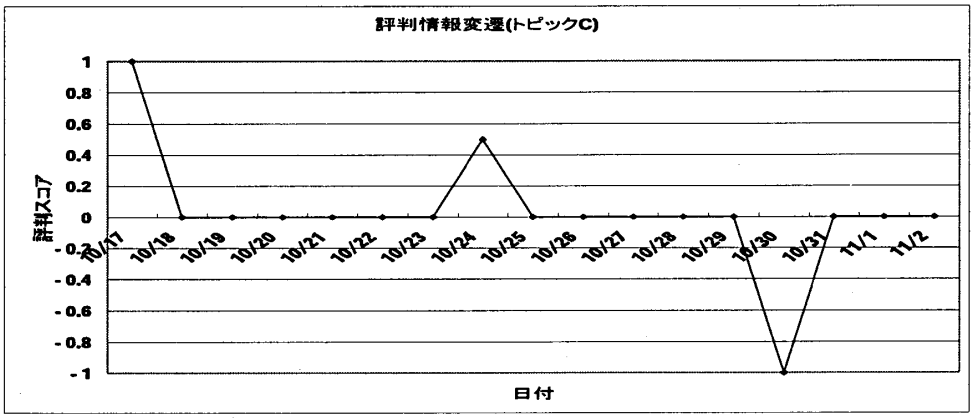

図 1　評判情報の変遷 (トピック C)

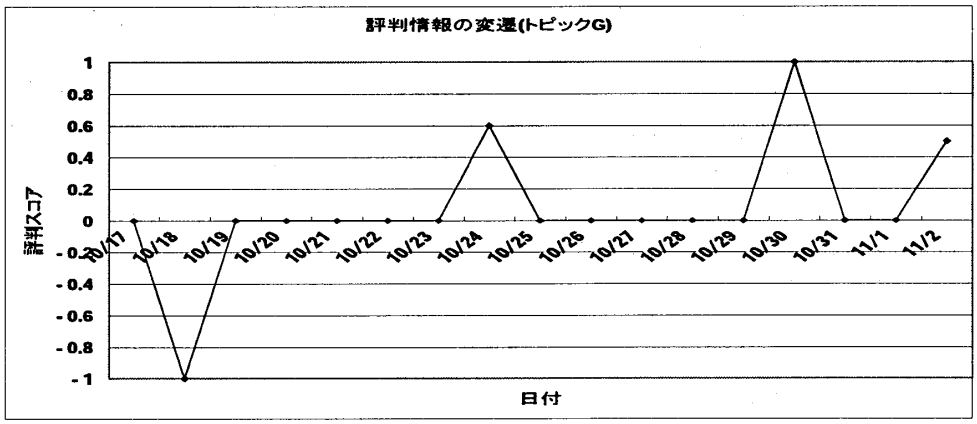

図 2　評判情報の変遷 (トピック G)

番号 19300026）の助成による．

参考文献

[1] 総務省，ブログ・SNS の現状分析及び将来予測，http://www.xoumu.go.jp/s-news/2005/pdf050517_3_1.pdf，2005.

[2] 戸田智子、福田直樹、石川博，Blog 記事のクラスタリングによるカテゴリ別話題変遷パタンの抽出，電子情報通信学会データ工学ワークショップ DEWS2007 A8-3，2007.

[3] kizasi.jp，http://kizasi.jp/

[4] Jon Kleinberg, Bursty and Hierarchical Structure in Streams, In Proc. the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[5] 藤木稔明、南野朋之、鈴木泰裕、奥村学，document stream における burst の発見，情報処理学会研究報告，2004-NL-160，pp.85-92.，2004.

[6] 立石健二、石黒義英、福島俊一，インターネットからの評判情報検索，情報処理学会研究報告，NL-144-11，pp.75-82，2001.

[7] 藤村繁、豊田正史、喜連川優，電子掲示板からの評価表現および評判情報の抽出，人工知能学会第 18 回全国大会，2004.

[8] 藤村繁、豊田正史、喜連川優，文の構造を考慮した評判抽出手法，電子情報通信学会データ工学ワークショップ DEWS2005 6C-i8，2005.

[9] 鈴木泰裕、高村大也、奥村学，Weblog を対象とした評価表現抽出，人工知能学会セマンティックウェブとオントロジー研究会，SIG-SWO-A401-02,2004.

[10] 霜田雄一、成田祐一，Blog からの評判抽出システムの構築に関する研究，平成 17 年度第 4 回情報処理学会東北支部研究会，2005.

[11] 形態素解析システム Sen，http://ultimania.org/sen/

[12] 石川博，次世代データベースとデータマイニング，第 6 章　クラスタリング，CQ 出版社，2005.