Twitterストリームのバーストの断続性に着目したキーワード抽出

坂本　翼† 廣田　雅春†† 横山　昌平††† 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
†† 静岡大学創造科学技術大学院　〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †gs11022@s.inf.shizuoka.ac.jp, ††dgs11538@s.inf.shizuoka.ac.jp ,
†††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

あらまし　Twitter の解析や要約においてキーワードの抽出は重要である．Twitter のあるひとつのトピックについて，記事のバーストの検出とそれに対応するキーワード抽出を行うことでそのトピックのトレンドを表わすことが可能である．あるトピックにおけるトレンドの変遷を追うためには，時間と共に変化するトレンドを表すようなキーワードを抽出する必要がある．本研究では，あるトピックの Twitter ストリームにおけるバーストの断続性に着目して，過去のバーストの情報を用いて新たなバーストを表わすようなキーワードを発見する手法を提案する．

キーワード　Twitter，バースト検出，キーワード抽出，要約

A Method of Keyword Extraction Focused on the Intermittent Bursts in Twitter Streams

Tsubasa SAKAMOTO†, Masaharu HIROTA††, Shohei YOKOYAMA†††, Naoki FUKUTA†††, and Hiroshi ISHIKAWA†††

† Graduate School of Information, Shizuoka University　3–5–1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
††† Department of Computer Science, Faculty of Informatics, Shizuoka University　3–5–1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
†† Graduate School of Science and Technology, Shizuoka University　3–5–1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
E-mail: †gs11022@s.inf.shizuoka.ac.jp, ††dgs11538@s.inf.shizuoka.ac.jp ,
†††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

Abstract　Keyword extraction from Twitter streams is important for summarizing twitter contents. Burst detection for keyword extraction from Twitter streams is used for discovering a trend of topics.. There is a method for extraction of keywords that shift with time, to trace the transition of the trends. In this paper, we propose a method to extract the keyword from streams and detect bursts using the information of prior bursts. Our approach focuses on the intermittent bursts in Twitter stream to extract keywords of new trends.

Key words　Twitter, Burst detection, Keyword extraction, Summarizing

1. はじめに

近年，Twitter [1] はマイクロブログサービスとして普及し，多くの人々に利用されている．Twitter では，1 億人以上のユーザが利用しており，そのツイートは 5 日当たり 50 億件以上になる(注1)．Twitter の投稿記事は 140 文字以内の短いテキストであり，投稿の手軽さから，利用者によってそのとき自分が見ているものや感じていることなどのリアルタイムな情報が記事として投稿される場合が多い．これらの記事は共通の話題に関心のあるユーザにとって，有用な情報源となっている．

(注1)：http://www.itmedia.co.jp/news/articles/1109/09/news027.html.

Twitter にはハッシュタグという機能があり，ユーザが記事を投稿する際に付与することで記事をグループ化することができる．ユーザはハッシュタグを用いて検索をすることで，興味のある話題に関する Twitter ストリームを閲覧することができる．

Twitter の特徴的な利用方法のひとつとして，テレビ番組の放送などある話題について，ある出来事の内容や，出来事に関した感想を投稿するために用いられることがある．多くのユーザがテレビを見ながら同時に投稿するため，Twitter にはその番組に関する非常に多くの記事が集まる．番組に関する投稿にはハッシュタグが付けられることが多く，ユーザはハッシュタグで検索することで，他の多くのユーザの番組に関する投稿記事を閲覧することができる．ハッシュタグが付与されている記事の多くはその番組で起きているイベントやユーザの感想などの情報を含んでおり，共通の番組の視聴者にとって有益な情報源となる他，番組を視聴しながら他のユーザと感覚を共有するといった使われ方もされている．

また，テレビ番組などに関する Twitter ストリームは，時間と共に Twitter ストリームに含まれる記事の内容が変遷するという特徴がある．内容の変遷を理解することで，テレビ番組を実際に見ていないユーザも番組の内容を理解することができる．しかしながら，ユーザが興味のある話題の情報を閲覧する際に，ハッシュタグのように記事がひとつの Twitter ストリームにまとめられていても，全ての記事を読むことは記事量が多いため困難である．そのため，本研究では，このような時間経過と共に内容が変遷するひとつの話題に関する Twitter ストリームを自動的に要約する手法の実現を目指す．本研究では，例えば紅白歌合戦のような，ひとつの話題をトピックと呼ぶこととする．あるトピックについての Twitter ストリームを要約する技術はユーザの閲覧の負担軽減という観点から重要である．

また，ユーザには進行中のトピックについて直近でどのような出来事が起きているのかを知りたいという要望もある．例えば，途中から視聴を始めた視聴者が番組の様子を知るという利用が考えられる．そのために，本研究では Twitter ストリーム動的に解析して自動的にトピックの要約を生成する技術の実現を目指す．

図 1 に 2011 年 12 月 31 日から 2012 年 1 月 6 日までの間に NHK のテレビ番組に関するハッシュタグ「#NHK」が付けられた Twitter の投稿記事数の 1 時間ごとの推移を示す．図 1 から，NHK 紅白歌合戦などの放送時間にユーザの投稿数が急増し，注目を集めていたことがわかる．これらのトピックに見られるような，ある時間において記事が集中する状態をバーストと呼ぶ．Twitter ストリームに対してバースト解析を行うことでユーザ達が注目したトピックを発見することが可能である．また，図 2 に紅白歌合戦の放送時間に「#NHK」が付けられた Twitter の投稿記事数の 1 分間ごとの推移を示す．図 2 では図 1 よりも詳細なバーストが断続的に発生しているという特徴がある．これはトピックの中にユーザが注目する出来事がいくつもあることを表している．このように，より詳細な時間でバースト解析をすることでトピック中の詳細なイベントを発見することが可能である．

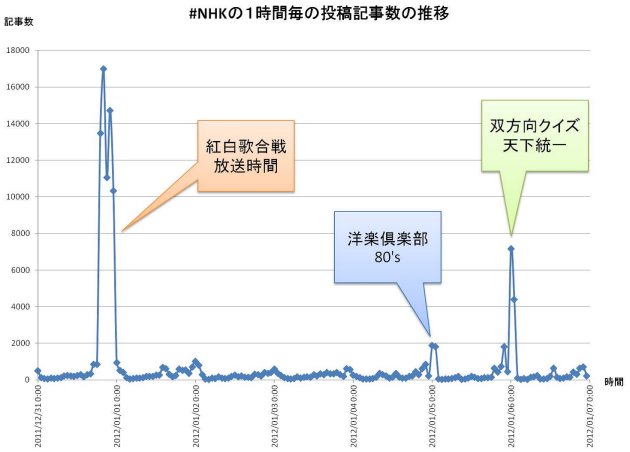

図 1　#NHK の 1 時間ごとの投稿記事数の推移

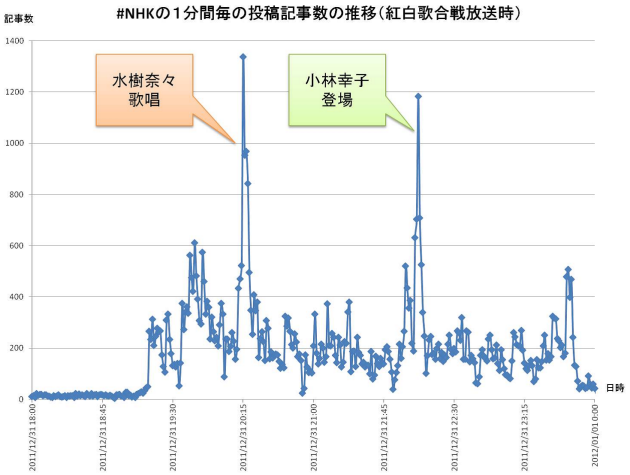

図 2　#NHK の 1 分間ごとの投稿記事数の推移

本研究では，トピックの内容を要約するため，トピック内で断続的に発生するイベントを利用する．イベントを検出するために，Twitter のストリームに対してバースト検出を適用し，バーストが発生している期間をイベントとする．このとき，各イベント期間中の記事をイベント関連記事として解析することで要約を生成し，順番に並べたものをトピックの要約とする．

我々は過去の研究において動的に Twitter ストリームの要約を行うシステムの試作を行なっている [2]．試作システムでは検出したイベントについて，代表記事を 1 つ選択することでイベントの要約として提示する手法を用いた．代表記事はイベント関連記事の中から他の記事との単語の被覆度を用いて決定する．この際，イベント関連記事をイベントの発生期間を用いて決定しているために，その中にはイベントとは関係のないノイズとなる記事も当然含まれる可能性がある．そのため，イベントにおける重要な単語を決定し，重要な単語のみについて単語の被覆度を計算することによってノイズ記事の影響を抑える手法を提案した．重要な単語の決定にはイベントにおける単語の出現数を指標とし，一定以上出現する単語を要約に必要な単語とした．しかし，出現数のみで重要な単語を決定すると，次の

ような不適切な単語が存在し，求める要約が得られない場合があった．

あるひとつのトピック内では，トピック全体を通して多く出現する単語が存在する．例えば，紅白歌合戦では「紅白」という単語がトピック中の多くの時間帯で出現回数が多い．このような単語をトピックにおける恒常的な単語と呼ぶこととする．イベントの内容を表すようなキーワードを抽出する際に，単語の出現回数は指標となるが，出現回数の多い単語をイベントのキーワードとした場合，恒常的な単語が出現回数の上位に来る事が多い．しかし,「紅白」等の多くの恒常的な単語は，イベントの内容を適切に表しておらず，イベントのキーワードとして適さないと思われる単語である．また，ユーザの投稿速度には差があり，イベントの発生から記事の投稿までには遅延が伴う．そのため，断続的にイベントが発生しているときに 1 つ前のイベントの内容に関する記事が新たなイベントの発生している時間に投稿されることがあり，ひとつのイベント発生期間の中に過去のイベントの情報が混在することになる．これらの理由から，ひとつのバースト内の記事情報だけからイベントのキーワードを正しく推定することは困難である．そこで，本論文では断続的にバーストが出現する場合に過去のバーストの情報を用いて重み付けをすることで，適切なキーワードを抽出する手法を提案する．

2. 関 連 研 究

Twitter の記事集合からキーワードを抽出する研究として Zhao らの研究 [4] がある．Zhao らは 1 つの記事は 1 つのトピックの内容を表すという仮説に基づく Twitter-LDA と呼ばれるモデルによって Twitter の記事集合をトピックごとに分類し，それらからトピックの内容を表すキーワードやキーフレーズを抽出している．本論文のキーワード抽出とは内容を同定する範囲がトピックと詳細な時間を表すイベントとで異なっている．また，本論文では共通のハッシュタグの付けられた記事は共通のトピックについての内容を表しているという仮定の基，ハッシュタグを用いることでひとつひとつの記事についてトピックの同定を行うことなく一つのトピックの解析を行なっている．

Twitter のトピックの要約を行った研究として，高村らの研究 [3] がある．高村らの手法では，トピックに関するエントリ（記事）の中から他のエントリとの単語の被覆度が大きいエントリを代表エントリとしていくつかのエントリを選択し，時系列順に並べることでトピックの要約を行っている．代表エントリはトピック中のイベントの要約として出力される．代表エントリを決定するため，トピックを時間軸上に並んだエントリの集合として捉え，施設配置問題を応用した要約モデルを提案している．高村らの手法ではトピックの要約を静的に生成することを目的としており，最大要約長となる代表エントリの数を指定する必要がある．本研究ではバースト解析によるイベント検出を行うためバースト解析のパラメータによって最大要約長は自動的に定まる．また，リアルタイムに発生する新たな Twitter ストリームを動的に要約することを目的としている点で異なる．

また，リアルタイムに Twitter ストリームに含まれる記事

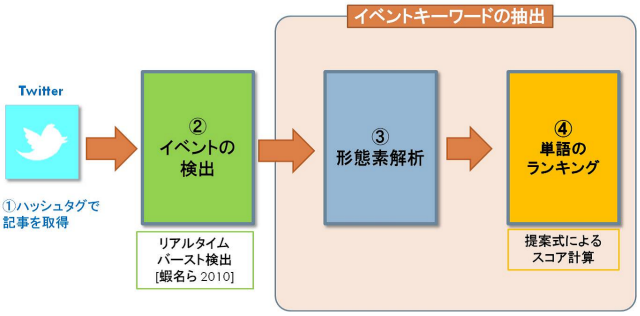

図 3　システムの概要

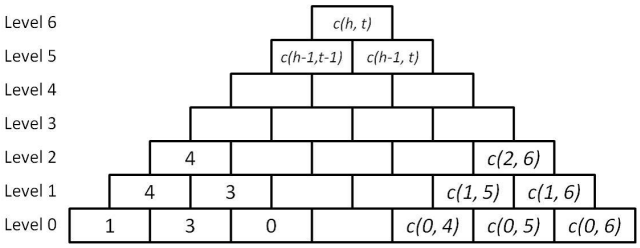

図 4　Aggregation Pyramid の例.

が変遷するトピックとして，国会討論などのディベートの解析を行った研究として Nicholas らの研究 [5] がある．Nicholas らは，まず Amazon Mechanical Turk によってテレビ番組でのディベートに関する Twitter の記事が positive か，negatie のどちらであるかの判定を人手で行い，精度の評価をしている．次に，positive，negative のバランスからディベートの特徴を解析し，投稿したユーザ達が興味を強く持った時間や論議が起こっている時間の検出を試みている．Nicholas らが提案するシステムは，記事を編集するジャーナリストや政治家が解析結果を利用することを想定している．

3. 提 案 手 法

図 3 に要約システムの試作の概要を示す．本手法では，トピックの要約を目的とし，ハッシュタグを用いて収集したテレビ番組などのトピックに関する Twitter の記事をバースト解析することで，トピック中のユーザーが注目するイベントを検出する．検出したイベントについて，イベントの要約を生成するためにイベントの内容を表すような単語群を決定する．

3.1 イベント検出

イベントの検出は時間当たりの投稿記事数を用いてバースト解析を行う．イベントの発生期間は検出したバーストの開始時間から終了時間までの期間とする．Twitter ストリームから動的にバーストを検出するために，蝦名らの提案したリアルタイムバースト検出手法 [4] を用いる．蝦名らの手法は従来のリアルタイムバーストの解析手法のように一定期間毎にバーストを解析するのではなく，ドキュメントの発生ごとにバーストを解析する．また，短時間に大量のドキュメントが発生した場合でも高速性を保つアルゴリズムを用いている．そのため，短期間に大量のドキュメントが発生する Twitter ストリームの解析に適していると考えられる．

蝦名らの提案したリアルタイムバーストの解析手法では

$AggregationPyramid$ と呼ばれるピラミッド状のセルデータ構造を用いる．図 4 に $AggregationPyramid$ の例を示す．ピラミッド構造のレベル 0 は N 個のセルを持ち，上層のセルは下層のセルの情報を統合したデータを持つ．各セルは記事の合計到着間隔 ($gaps$)，到着時間 ($arrt$)，間隔個数 ($gapn$) のデータを持つ．バーストの判定は到着間隔を指標として行う．到着時間が重複していない直前の状態と比較して発生間隔が急激に小さくなっている期間をバーストと判定する．各セルについて平均到着間隔関数を式 (1) のように定める．

$$avg(c(h,t)) = \frac{c(h,t).gaps}{c(h,t).gapn} \tag{1}$$

セルの平均到着間隔 $avg(c(h,t))$ と直前の状態の平均到着間隔 $avg(c(N-1,t-1-h))$ を比較し，各レベルの最新のセルについてバーストを解析することで複数のウィンドウサイズに渡ったバーストを判定することができる．バースト検出のパラメータとして，ピラミッド構造のサイズ N，バースト係数 β，セルの最小ウィンドウサイズ W_{min}，バースト判定を行う最低の記事数 $Amin$ を設定する．

この手法によって Twitter の記事が到着するたびにバーストをしているか判定を行い，イベントを検出する．

3.2 キーワードの抽出

検出したイベントについてその内容を表すキーワードを抽出する．キーワードを抽出するために，まずイベント中に出現する各記事を形態素解析する．記事の形態素解析には MeCab [7] を用いる．各記事を MeCab の形態素解析によって解析した結果の単語群から内容語 (名詞・動詞・形容詞) を抽出して登録する．このとき，形態素解析の結果として得られた単語の中から非自立語，接尾語，および代名詞をストップワードとする．また,「する」，および「なる」などの文章の特徴となりづらい単語や，記号のみで構成される単語，および平仮名・カタカナ一字の単語もストップワードとする

記事からキーワードを抽出するために，単語の重要度を決定するスコアを計算し，イベントにおける単語のランク付けを行う．同じトピック中で複数のイベントが存在するとき，これらのイベントではトピック内で共通の単語が一定量出現するという特徴がある．また，ユーザが投稿を完了するまでの時間にはばらつきがあり，直前に発生したイベントに関する内容の記事が新たなイベント中に一定量出現するという場合がある．これらの単語は新たなイベントの内容を表すキーワードとしては不適切であると考えられるため，これらのスコアを抑える重み付けを行う．

ここで，トピック中に発生した n 番目のイベントに含まれる記事の集合 D_n を $D_n = \{d_n^1, d_n^2, \cdots, d_n^l\}$ と定義する．このとき，n 番目のイベントに含まれる記事の総数は $|D_n|$ となる．また，D_n に含まれる単語の集合 W_n を $W_n = \{w_n^1, w_n^2, \cdots, w_n^m\}$ と定義する．また，D_n 内において出現するある単語 w_n^i を含む記事の総数を $DocNum(n, w_n^i)$ とする．

イベントの内容をよく表すようなキーワードを含む記事は，イベント内の記事の中でで大きな割合を占める．そのため，新たなバーストの内容をよく表すようなキーワードを含む記事は，直前のイベントと比較して，バースト内の記事総数に占める割合が増加していると考えられる．また，直前のイベントで注目された内容に関する記事がユーザーの投稿速度の遅延によって新たなイベントでそこで，単語 w_n^i のスコアの要素として，ひとつ前のバースト内で単語 w_n^i が含まれる記事の割合と，新たに発生したバースト内で単語 w_n^i が含まれる記事の割合の増減率 $PercentRate(n, w_n^i)$ を用いる．このとき，$PercentRate(n, w_n^i)$ は以下の式 (2) で表される．

$$PercentRate(n, w_n^i) = \frac{DocNum(n, w_n^i)}{|D_n|} \Big/ \frac{DocNum(n-1, w_n^i)}{|D_{n-1}|} \tag{2}$$

イベントにおける単語 w_n^i を含む記事数に重みとして式 (2) を組み合わせた式 (3) を n 番目のバーストにおける単語 w_n^i のスコア $Score(n, w_n^i)$ として定義する．

$$Score(n, w_n^i) = DocNum(n, w_n^i) \times PercentRate(n, w_n^i) \tag{3}$$

ここでは，$n = 1$ のときのスコアは $DocNum(n, w_n^i)$ とした．また，$DocNum(n-1, w_n^i) = 0$ のとき，$PercentRate(n, w_n^i)$ は以下の式 (4) とした．

$$PercentRate(n, w_n^i) = \frac{DocNum(n, w_n^i)}{|D_n|} \Big/ \frac{1}{|D_{n-1}| + 1} \tag{4}$$

これによって，直前のイベントで出現している単語に対して全く出現しなかった単語は僅かに優位に働く．上記の式 (3) を用いて単語のスコアを計算し，ランキングを生成する．

4. 評価実験

提案手法の有効性を実証するための評価実験を行う．評価実験では提案手法を用いてイベントにおける単語の重要度によるランキングを生成し，実際に重要だと判断される単語がどれだけ上位に存在するか，不要な単語が上位に存在しないかを評価する．

まず，複数のトピックを含む記事データを用いてバースト解析を行い，イベント検出を行う．バースト解析のパラメータは経験的に $N = 60$，$\beta = 0.80$，$W_{min} = 3000$(ミリ秒) を採用した．また，$Amin$ はバースト判定時に (判定を行うセルのレベル)× 5 とした．さらに，イベントとするのは検出したバーストのうち発生期間が 15000 ミリ秒以上のものとした．

イベント検出の際，Twitter のリツイート記事は解析の対象外とした．リツイート記事は他の記事を引用した記事であり，同じ内容が何度も出現するため一部の単語の出現数が大きくなる．また，引用されるために時間的に隔たりを持って出現するという性質上，イベント内容を表すキーワードの抽出においてノイズとなりやすい．tuneTV(注2)による自動ツイートを含む記

(注2)：http://itunes.apple.com/jp/app/id448518322?mt=8. Twitter 連動型ソーシャルテレビアプリ．番組を視聴していることを定型文を連動しているソーシャルサービスに投稿して知らせるチェックイン機能がある

事も同様の理由で一部の単語の出現数が大きくなりノイズとなるため解析の対象外とした．検出したイベントのうち実験に用いるデータとして，3 つのトピックからそれぞれ 10 個のイベントをランダムに選出した．

次に，各イベントにおける内容を構成する単語群の正解データを人手で作成した．各イベント期間中の記事が含む単語について，それぞれの単語がバースト期間が示す記事内容のキーワードとして適しているかどうかを実験協力者によって人手で判定し，各イベントの正解となるキーワードの集合を作成した．判定したのはバースト期間中の単語のうち各単語を含む記事の出現数が上位３０件の単語である．まず，実験協力者はイベント期間中の記事を読んで記事を投稿しているユーザがどのような内容に注目しているのかを把握し，次に各単語がその内容に対してノイズであるかどうかを判定した．この際，実験協力者には恒常的な単語に注意をするように指示した．

評価実験では各イベントについて，提案手法によるキーワードのランキングと比較手法によるキーワードのランキングによってそれぞれバースト期間内に出現する単語のランク付けを行い，上位 30 語について比較する．比較手法では単語を含む記事の出現数によるランク付けを用いる．

4.1 実験データ

解析対象とするデータセットとして，テレビ番組に関するハッシュタグを含む以下の 3 つのデータセットをハッシュタグクラウド [8] より取得した．

(1) #NHK を含む 2011/12/31 の記事
(2) #NHK を含む 2012/1/6 の記事
(3) #laputa を含む 2011/12/9 の記事

これらのデータセットはそれぞれ以下の TV 番組の放送時間をを含む．

(1) NHK「第 62 回 NHK 紅白歌合戦」
(2) NHK「双方向クイズ 天下統一」
(3) NTV「金曜ロードショー『天空の城ラピュタ』」

これらのデータセットからバーストを検出した結果，それぞれのテレビ番組に関してそれぞれ (1)52 個，(2)17 個，(3)35 個のイベントを検出した．これらのイベントからそれぞれのトピックについて 10 個ずつをランダムに選出し，評価実験を行った．

4.2 評価指標

評価指標としてランキング評価指標である $nDCG@K$ を用いる．$nDCG@K$ はランキングの上位 K 件について，理想的なランキングへの近さを表す．$nDCG@K$ の評価値は以下の式 (5) で表される．

$$nDCG@K = \frac{1}{IDCG} \sum_{j=1}^{K} \frac{(2^{r(j)} - 1)}{log(1 + j)} \qquad (5)$$

$IDCG$ は理想的なランキングを表し，上位 K 件の単語が全て正解である場合の評価値である．$r(j)$ はそれぞれのランク j の評価値を表し，上位 j 番目の単語が正解ならば 1，不正解ならば 0 とした．

4.3 評価結果

図 5 に実験対象とした 30 イベントにおける $nDCG$ の評価値の平均値の推移を示す．$K = 6$ 以降は僅かに提案手法の評価値が低い結果となった．詳細な評価値を見るために表 1 に上位 10 件の平均評価値を示す．上位 5 件では比較手法と提案手法の有効性が入れ替わっているが 6 位以降では提案手法は評価値が低い結果となっている．

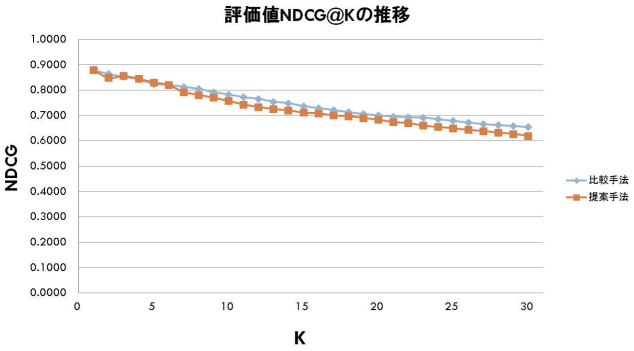

図 5　nDCG@K の平均値推移

表 1　上位 10 件の平均評価値

K	比較手法	提案手法
1	0.8788	0.8788
2	0.8636	0.8485
3	0.8527	0.8558
4	0.8424	0.8449
5	0.8248	0.8307
6	0.8212	0.8206
7	0.8134	0.7928
8	0.8050	0.7815
9	0.7924	0.7704
10	0.7834	0.7575

表 2　上位のキーワードの例

	比較手法	提案手法
1	ミッキー	ミッキー
2	嵐	嵐
3	ディズニー	ドナルド
4	紅白	ディズニー
5	いい	中
6	くる	ダンス
7	中	キレ
8	人	チップ
9	福	ぐるみ
10	ドナルド	着

原因として提案手法の評価の平均は有利に働く場合と不利に働く場合が存在した．表 2 にイベントの上位 10 件のキーワードの例を示す．イベントは紅白歌合戦の番組に関するものであり，内容はディズニーキャラクターが登場して歌手グループである嵐と共演してパフォーマンスを行ったものである．比較手法である単純な出現記事数によるランキングでは 4 番目に紅白歌合戦の番組において恒常的な単語である「紅白」が出現しているが，提案手法では他のイベントでも一定の割合出現しているためにスコアが下がり，ランクを下げることに成功している．また，比較手法の 9 番目の「福」は直前のイベントで出演した

「鈴木福」を示すが，個々のイベントで大きく注目されている内容とは異なる．これに対して，提案手法ではランクを下げることに成功している．一方で，比較手法の上位に含まれる「いい」「くる」「人」は正解として選択されていた．しかし，これらの単語は一般語であり直前のイベントでも出現しており，提案手法ではスコアを大きく下げてしまっていた．提案手法は特徴的なキーワードに対しては効果があったと考えられるが，一般語に対しては別の手法を検討する必要があると考えられる．

また，提案手法の上位に「ぐるみ」「着」という単語があるが，これは「着ぐるみ」の形態素解析ミスである．「着ぐるみ」はこのイベントについて内容を表したものであったが，形態素解析ミスのために正解として選択されていなかった．このため，N-gram などによって形態素解析ミスを補完することは精度の向上に有効だと考えられる．その他，「ディズニー」「でぃずにー」のように表記揺れによって出現数が少なくなってしまう場合があるため，同じ意味の単語を推定する手法も精度の向上に役立つと考えられる．

また，複数のバーストにおいて同一の人物が注目されているような場合はそれらのどのバーストにおいてもその人物名を表す単語のスコアが高いことが望ましい．しかし，提案手法では前のバーストにおいて出現頻度が高い場合に次のバーストでスコアが他の単語と比較して低くなってしまう．提案手法ではひとつ前のバーストの情報のみを利用しており，より過去のバーストの情報も用いることや，単語の共起関係をスコアに組み込むこと，バースト間の単語の類似度などからバースト同士の関連度合いを求めて単語の重みを変えることで精度を高める手法などが考えられる．

5. ま と め

本論文では，トピック中に起こるイベントの要約を生成することを目的とし，内容を表すようなキーワードを抽出することを課題とした．イベントの内容を表すようなキーワードは単純な出現数にはよらず，出現数の上位にはトピックで恒常的に出現する単語や，ユーザの投稿時間の違いによって出現する単語が存在する．そこで，Twitter ストリームの要約を目的として，トピック中に起こるイベントの内容を表すようなキーワードを過去の単語の出現情報を用いて抽出する手法を提案した．また，Twitter ストリームの要約のために，トピックの内容はトピック中で断続的に発生するイベントによって構成される，というモデルを示した．イベントの検出にはトピックの Twitter ストリームの記事数を対象としたリアルタイムバースト検出手法を用い， 実験によってトピック内のイベントを検出できることを確認した．イベントの内容に関するキーワード抽出では，トピックにおけるバーストの断続性に着目し，直前のバーストの情報を用いて単語に重み付けをすることで，新たなイベントの内容を表すようなイベントのキーワードを抽出する手法を提案した．また，実験の結果から本手法が有効に働く場合と，不適切に働く場合があることを確認した．実験の結果をもとに提案手法のスコアの計算式を改良し，課題の解決を図る．

今後はより多くのテレビ番組にも適用する他，テレビ番組以外のトピックに関しても適用し，手法の有効性について調査する．

文 献

[1] Twitter，http://twitter.com/.

[2] 坂本翼，横山昌平，福田直樹，石川博．マイクロブログを対象としたリアルタイムな要約生成システムの試作．The 3rd Forum on Data Engineering and Information Management.

[3] Hiroya Takamura，Hikaru Yokono and Manabu Okumura. Summarizing microblog stream．人工知能学会研究会資料 SIG-SWO-A1001-03.

[4] Zhao，X. and Jiang，J. and He，J. and Song，Y. and Achananuparp，P. and LIM，E.P. and Li，X. Topical keyphrase extraction from Twitter．The 49th Annual Meeting of the Association for Computational Linguistics.

[5] Diakopoulos，N. A. and Shamma，D. A. Characterizing debate performance via aggregated twitter sentiment. Proceedings of the 28th international conference on Human factors in computing systems.

[6] 蝦名亮平，中村健二，小柳滋．リアルタイムバースト検出手法の提案．日本データベース学会論文誌，Vol.9，No.2 November 2010.

[7] MeCab，http://mecab.sourceforge.net/.

[8] ハッシュタグクラウド，http://hashtagcloud.net/.