Twitter解析のための技術と2013年ヒット予測

山川　義介，小野　広司

日経トレンディは，年末にその年の「ヒット商品ベスト30」と，翌年の「ヒット予測ランキング」を発表している．本研究では，2013年のヒット予測ランキングを，その商品に関連するワードのTwitter解析により行う．Twitter解析に関しては，その活用の可能性，データ収集方法，解析の際に利用できる技術についても紹介する．また連鎖的に盛り上がるバースト現象に関して，その定性的要因についても言及する．

キーワード：Twitter，テキストマイニング，API，バースト解析

1. はじめに

Twitterは，2006年に米国でサービスインした140文字以内の短い文章を投稿できるSNSで，ミニブログとも言われる．Facebookやmixiと違い，フォローをすれば相手が承認しなくても投稿を読むことができる．他のユーザーの投稿を引用形式で自分のアカウントから発信するリツイート機能を使うと，連鎖的に話題が盛り上がることもあり，時代のトレンドを見る一大メディアとしての存在感も増している．

一方，日経トレンディ[1]は，毎年その年のヒット商品と，翌年のヒット予測ランキングを年末に掲載している．No.345（2012年12月号）によると，2013年のヒット予測ランキングは以下のとおりとなっている．

順位	商品名
1	日本流ロングトレイル
2	「抗ロコモ」ギア＆フード
3	手のひらタブレット
4	新・希少糖ドリンク
5	でぶ犬予防サービス
6	ななつ星／サン・プリンセス
7	七変化ウォーターフレーバー
8	イタリア・アート
9	ノンアル珍味
10	“ブランド香水”ミニ柔軟剤
11	スマホ風機能追加デジカメ
12	デジタル対応上質ノート
13	個人間スキル取引
14	サブクリ通販
15	シナプソロジー
16	アースカフェ
17	ネオ裏原系
18	マイクロEV
19	ポケットサイズ燃料電池
20	伊勢・出雲詣で

本研究では，これらの商品に加えALBERT分析チームで抽出したヒット予測商品に関連するワードを選定し，そのワードが2013年1月1日から4月30日の間にツイートされたデータを取得・分析し，2013年のヒット商品が何になるかの予測をした．

関連ワードは，記事内に記載されていたワードや，商品を検索し話題になっていたワードを独自に収集した．一例を以下に示す．

1. 日本流ロングトレイル
   - スーパートレイル
   - とかちロングトレイル
   - トレラン
   - ハイキング
   - ロングトレイル
   - 塩の道トレイル
   - 高島トレイル
   - 信越トレイル
   - 登山
   - 八ヶ岳山麓スーパートレイル
   - 歩く旅
   - 北根室ランチウェイ
2. 「抗ロコモ」ギア＆フード
   - NIKE + FuelBand
   - ロコトレ
   - ロコフィットGL
   - ロコモ

やまかわ　よしすけ，おの　こうじ
株式会社ALBERT
〒151–0053 東京都渋谷区代々木2–22–17

- ロコモディブシンドローム対策

3. 手のひらタブレット
   - 7 インチタブレット
   - GALAPAGOS A01SH
   - iPad mini
   - kindle fire
   - Nexus7
   - Xperia Tablet
   - アクオスパッド
   - メディアスタブ
   - レグザタブレット AT570
   - 手のひらタブレット
   - 小型タブレット
   - 中華 pad

2. Twitter 解析の技術

本章では Twitter 解析の対象となるデータの特徴を確認し，Twitter 解析がどのような場面で活用可能であるかについて紹介する．さらに，実際にツイートデータを収集する方法，および，解析の際に利用できる技術について紹介する．

2.1 Twitter 解析で扱うデータの特徴

具体的に Twitter 解析の活用について紹介する前に，解析の対象となるデータ，すなわち Twitter ストリーム上を流れるデータの特徴について確認しておきたい．

まず一つ目の特徴として，世間一般の興味関心度合いが表出されやすいという点が挙げられる．ある特定のトピックがあった際，そのトピックに対する Twitter ストリーム上での興味関心度合いは，そのトピックに関するツイート数という形で表出する．Twitter では投稿自体も手軽に行えるうえに，リツイート機能によりツイートの拡散が可能であるため，注目度の高いトピックに関するツイート数は，そのトピックの発生から短期間の間に際立って増加しやすい傾向がある．

二つ目の特徴として，データ量が膨大であるという点が挙げられる．Twitter ストリーム上には毎秒数千の単位で新規ツイートが生み出されている．このため Twitter 解析のシステムを構築する際には，数理的な解析ロジックを考えるのみならず，効率的に解析処理を行うためのデータ構造，アルゴリズムについても考慮する必要がある．

三つ目の特徴として，本文が自然言語を用いて記述されているという点が挙げられる．Twitter ストリーム上のデータを解析してその結果を利用する際には，データの中に表記揺れや表示上不適切な単語を含んでいる可能性があることに留意する必要がある．

2.2 Twitter 解析の活用

Twitter 解析の活用分野としてまず挙げられるのはマーケティング領域である．株式会社ホットリンクが提供するソーシャルメディア分析ツール「クチコミ@係長」は Twitter データの分析にも対応し，キャンペーンの効果測定や製品の市場調査を行うことができる [2]．

次に，バースト判定システムへの活用が挙げられる．バーストとはある期間の時系列データにおいて，ある話題が急激に増加する現象のことを言う．前節で述べたように Twitter は拡散性の高いソーシャルメディアであるため，多くのユーザーの興味をひくトピックがあった場合，その影響が Twitter ストリーム上に反映されやすい．したがって，ツイートの時系列データはバースト判定の対象として適していると言える．

バースト判定システムは，本や映画，音楽などのコンテンツ販売サイトにおいて，話題になっている作品，著者などの情報をエンドユーザーに提供するという活用が考えられる．あらかじめ判定対象とする作品，著者などのキーワードリストを作成しておき，Twitter ストリーム上においてリスト内のキーワードを含むツイート数を時系列に観測することで，現在話題となっているキーワードを判定することができる．このようなシステムでは情報の鮮度が重要となってくるため，リアルタイムでの解析が望ましい．

さらに，著者らは，商品に関連するツイート数の情報からヒット商品を予測する技術についても研究を行なっている．本稿では，本技術による予測結果の一例を紹介する．

2.3 ツイートの収集

Twitter 解析を行うためには，解析の目的に合ったツイートデータの収集をなう必要がある．Twitter サービス本体では，サービス開始当初のツイートから新たに生み出されるツイートまですべてのデータを保持していると考えられるが，これらすべてのツイートにアクセスできる API は一般には提供されていない．また，“Rate Limits” と呼ばれる 1 アカウントごとの API アクセス回数の制限も存在し [3]，解析システムを設計・運用する際には，これらの点に留意する必要がある．

以降，Twitter 解析の技術開発を行う過程で著者らがツイート収集に用いた API として，Twitter 社が提供する “REST API”[4]，Topsy 社が提供する “Otter API”[5] の一部の機能について紹介する．

(1) REST API–Search（以降，Search API と表記）

Search API では指定条件に合致するツイートを取得することができる．本文に特定のキーワードを含むツイートを取得したい場合などは本 API の利用が適している．ただし 1 回の検索で取得可能なツイート件数は 100 件までであり，検索時点から 6〜9 日以上前のツイートは取得できない [6].

(2) Streaming API

Streaming API は Twitter ストリーム上に時々刻々と生成されるツイートをリアルタイムにプッシュ配信する API である．エンドポイントとしてはすべてのツイートをプッシュ配信する “statuses/firehose” や，すべてのツイートの約 1%をランダムにサンプリングして配信する “statuses/sample” などが存在する．ただし “statuses/firehose” は Twitter 社とのライセンス契約が必要であり，一般には利用が許可されていない.

Streaming API は，辞書に登録しておいたキーワードが Twitter ストリーム上で出現する頻度（または，出現頻度のキーワード間における相対的な割合）を知りたい場合などに利用できる.

(3) Otter API–Search （以降，Otter API と表記）

Twitter 社が提供する Search API では直近 1 週間程のツイートしか取得できないのに対し，Otter API では過去長期間遡ってツイートデータを取得することができる．本稿で紹介する“ヒット商品予測”においても，検証にあたって予測対象の商品を連想させるキーワードを含む 2013 年 1 月 1 日以降のツイートが必要であったこと，また，あとからキーワードが追加される可能性があったため，本 API を用いてツイートデータの収集を行っている.

Twitter 社が提供する Search API と同様，検索条件としてキーワードを指定可能であるが，ツイート本文内に指定キーワードが含まれるツイートのみではなく，ある程度キーワードに類似した文字列を含むツイートも検索にヒットすることに留意して利用する必要がある．明確な検索仕様は公開されていない．なお，後述するヒット商品予測においては，API で取得したデータに対して再度検索語によるフィルタリングを実施することで，キーワードを含まないツイートを除去している.

また，著者らが本 API を利用してツイートデータを取得した際．レスポンスにおいてツイート本文が格納されるべき部分が空文字で返却されることがあり，一部動作が不安定なことがあった．利用に際しては，あらかじめ自身で API を試して確認してみることを推奨する.

2.4 Twitter 解析で利用できる技術

本節では Twitter 解析システムの実現に際して利用できる技術として，キーワードを含むツイートの効率的な検出方法について紹介する.

この方法は Twitter 解析において，キーワード集合内の要素を含むツイートをリアルタイムに検出したい場合に役立つ．Search API を用いて定期的に検索を行えば目的のデータは取得可能であるが，キーワード数が大きくなった場合，すべてのキーワードに対して定期的な検索を行うのは Rate Limits の制限上好ましくない．そのような場合，Streaming API でツイートを取得したツイートに対してキーワードのマッチングを行うことで，キーワードを含むツイートを検出するという方針が考えられる．（“statuses/sample” を用いる場合は全体の 1%のツイートしか検出できないが，キーワードを含む全てのツイートの取得が必要であれば，Streaming API での検出をトリガーとして Search API での検索を行えばよい．これにより，Twitter ストリーム上であまり出現しないキーワードに対する検索回数を抑えることができる.）

ここで，あらかじめキーワード集合が与えられている条件下で Streaming API によってツイートが 1 つ取得されたとき，そのツイート本文がキーワード集合内におけるどのキーワードを含んでいるかを検出する問題を考える.

最も単純な方法としては，各キーワードを順番にマッチングをしていく方法が挙げられるが，この方法では，マッチング判定処理の回数がキーワード数に比例することになる．Streaming API において秒間に取得されるツイート数を考えると，キーワードの数が多い場合にはこの方法は得策ではない.

そこでキーワード集合を Trie 木というデータ構造で表現することにより，効率的な検出を行うことができる．図 1 はキーワード集合を Trie で表現した例である．図中において網掛け表記のノードは，そのノードがキーワードの終端であることを示す.

図のような Trie 木があったときに，「dbcca」という文字列にどのキーワードが含まれているかを判定する場合を考える．図中の探索ステップにあるように，対象の文字列を 1 文字ずつに分解し，Trie 木を辿ることで検出を行う．各文字に対する処理においては，探索始点の子ノードを辿り，その子が当該文字と一致した場合，次の文字における探索始点にその子ノードを追加する．また，キーワードの終端ノードに達した場合

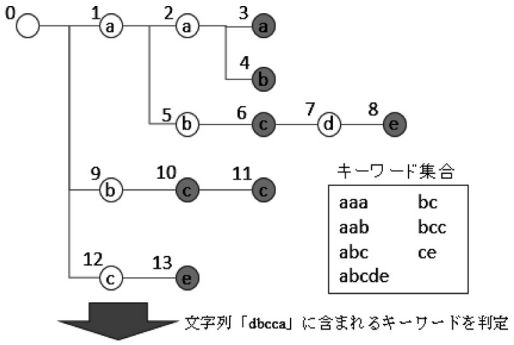
にキーワードを含むと判定され，そのノードから木の根に向かって文字列を辿ることにより，当該キーワードを復元できる．このようなステップでキーワードの検出を行ったとき，「dbcca」という文字列には「bc」「bcc」というキーワードが含まれていることがわかる．

なお，上記のような文字列処理のみで単語検出を行うと，例えば「私の出身地は東京都です」といった文があった場合，「京都」というキーワードも検出されてしまう．この問題を回避するためには，あらかじめ分かち書きしておく必要がある．形態素解析を行うと上記の例文において単語境界は「私 | の | 出身地 | は | 東京都 | です」となる．ここで，上記した探索アルゴリズムに下記ルールを追加することで，上記の問題に対応することができる．

- 現在の文字が単語の 1 文字目でない場合は Trie 木の根からの探索は行わない．
- Trie 木の終端に達した際，その文字が単語の最終文字でない場合は検出文字列としない．

著者らは約 1,000 単語のキーワード集合から生成した Trie 木を用い．Streaming API (statuses/sample) で取得したツイートに対してリアルタイムにキーワード検出処理を行ったが，問題なく処理できることを確認している．

また，Trie のデータ構造をさらに発展させて計算量を削減する Aho-Corasick 法という手法も存在する [7]．



文字	探索始点	一致	不一致	検出文字列
d	0	なし	1,9,12	-
b	0	9	1,12	-
c	0,9	12,10	1,9	bc
c	0,12,10	12,11	1,9,13	bcc
a	0,12,11	1	9,12,13	-

図 1　Trie のデータ構造と探索方法

3.　分析結果

3.1　バースト分析

図 2 に日経トレンディ予測上位 3 位の商品についてのツイート数を示した．各ツイートは特定日に突出してツイートが増加していることが見てとれる．Kleinberg のバースト解析 [8] では，特定のワードのバーストの様子や，バーストかどうかの判定，バースト度を用いたキーワードのランクづけをすることができるが，本研究では各商品に見られたバーストが，どのような要因で起きていたかを元のツイートに遡って分析を行った．

バースト理由としては，圧倒的にニュースや特集，記事などメディアによって発信されたものが原因になっているものが多く，イベントの開催なども挙げられる．Twitter の特性上，以下のようにニュースをリツイートするケースが多く，さらにそのリツイートを読んで

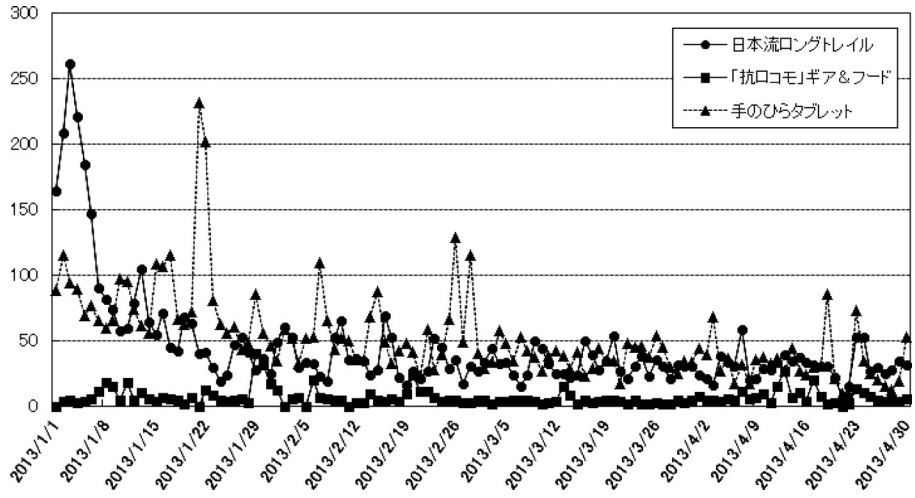


図 2　トップ 3 商品関連ワードのツイート数

表 1　バーストが見られた商品のバースト理由

商品名	バースト日*	バースト理由
日本流ロングトレイル	2013/1/3	「ハイキングやウォーキングは創造性を高めるという研究結果」という記事がライフハッカー[日本版]から出ていた(40件)amazonから「改訂版 ～県の山（新・分県登山ガイド）」という本が出版され、多くツイートされていた(25件)また、傾向としてお正月に高尾山や富士山に登る方が多かった
「抗ロコモ」ギア＆フード	2013/1/29	テレビ番組のミヤネ屋にて、メタボとロコモが特集されたため
手のひらタブレット	2013/1/21	SONYから「Xperia Tablet Z」のニュースリリースが出る
新・希少糖ドリンク	2013/1/4	アサヒ飲料のトクホ炭酸「ファイバー7500」が発売1か月で720万本発売されたことがニュースになる
ななつ星／サン・プリンセス	2013/2/4	Yahoo!ニュース、YOMIURI ONLINEにて特集される
イタリア・アート	2013/3/2	上野国立西洋美術館にて、「ラファエロ展」が開催される
“ブランド香水”ミニ柔軟剤	2013/4/25	メルサボンから、アーティストの「ゆず」とコラボレーションしたゆずの香りの新商品が発売された
スマホ風機能追加デジカメ	2013/1/29	キヤノンから新製品「PowerShot N」が発表される
デジタル対応上質ノート	2013/1/9	テレビ番組のヒルナンデスにて、スマレコノートが取り上げられたため
個人間スキル取引	2013/1/7	株式会社nanapiの社長古川健介氏が書いた「僕がユーザーとして本当に感動したWebサービスまとめ(2013年版)」内で取り上げられる
ポケットサイズ燃料電池	2013/1/8	「2週間持つモバイル“燃料”バッテリー「Nectar」が発表、夏にも発売開始」のニュースが出回る
ソーシャルゲーム	2013/1/6	12月28日～1月6日までパズドラ内でイベントが開催されており、その最終日にレアなキャラが発生したりした様子
新規商業施設	2013/3/21	東京駅前KITTEオープン日
料理系サイト	2013/1/10	9日からクックパッドのプレミアサービス1年分があたるスピードくじが行われる
スマート～	2013/2/7	エンジニアtypeに掲載された「現役大学生ら3人が開発した、「普通の家電」をスマート家電にするデバイス『Pluto』がすごい【連載:NEOジェネ！】」という記事が出回る(92件)
次世代テレビ	2013/1/8	SONYから「世界初／世界最大“56型4K対応有機ELテレビ”を開発」のニュースリリースが出る

*期間中の最大ツイート数を記録した日を「バースト日」とした

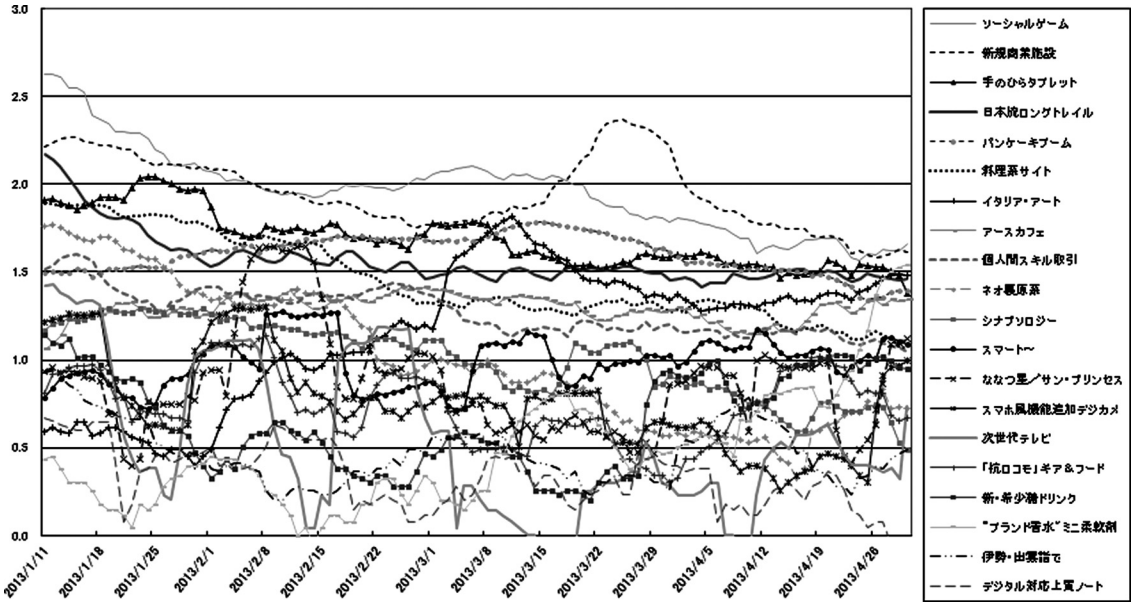


図 3　全商品のツイート時系列推移

リツイートするという連鎖反応によって，急激にその話題に関するツイートが増加すると考えられる．

> 「これ，Wi-Fi 版も出てくれないかなぁ… Xperia Tablet Z は iPad よりも薄くて軽いのに防水仕様 週アス PLUS #asciiplus http://weekly.ascii.jp/elem/000/000/124/124512/ . . . 」

「Twitter@tenin_sato 2013 年 1 月 21 日」より引用

3.2　トレンド分析

次に，各ワードが時系列でどのように変化をしているかを分析した．前記のとおり，ツイートデータにはバースト現象があり，中長期の予測をするにはそのままのデータでは判断しにくいことから，10 日移動平均をとり，さらに対数をとったグラフを図 3 に示した．新規商業施設の関連ワード上位ツイートは，ほぼ KITTE と HIKARIE で占めているが，3 月 15 日の東急東横線の地下化，3 月 21 日の KITTE オープンが重なった

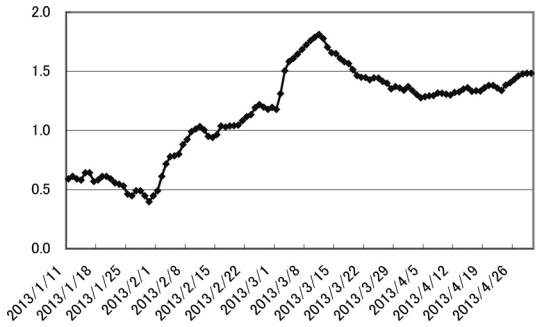


図 4　イタリア・アートのツイート推移

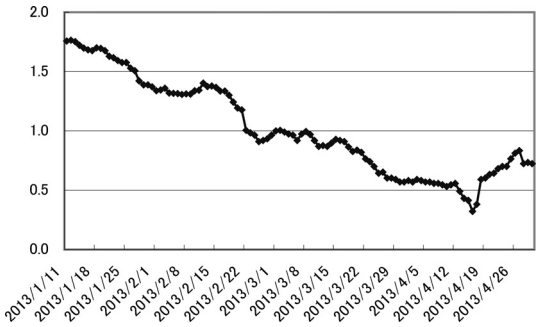


図 5　ネオ裏原系のツイート推移

ことで，新規商業施設に関連するツイートが数日にわたりバーストしたことがわかる．

さらに個別に見てみると，この 4 カ月で急上昇している関連ワードがあることがわかる．ラファエロ，ダ・ヴィンチ，ミケランジェロなど，今年は前代未聞のイタリア芸術の当たり年だそうで，「イタリア・アート」は，上野国立西洋美術館にて「ラファエロ展」が開催されたことで，図 4 のように後半の 2 カ月で一気にツイートが増加した．

逆に，再ブレイクするかと思われた，90 年代に一世を風靡した厚底，ダメージデニム，ニットキャップなどの「裏原系ファッション」は，図 5 のように，年始以来どんどん話題に上らなくなっている．

3.3　総合順位

最終的な 2013 年ヒット商品予測は，1 月から 4 月までの関連ワードツイート総数に，ALBERT 独自の時系列予測モデルであるトレンド係数を加味し予測を行った．結果を次に示す．

順位　商品名

1　※新規商業施設（KITTE，HIKARIE など）
2　※ソーシャルゲーム（にゃんこ大戦争，パズドラ，ラブライブなど）
3　イタリア・アート
4　※パンケーキブーム（エッグスシングス，カフェカイラ，パンケーキデイズなど）
5　手のひらタブレット
6　日本流ロングトレイル
7　アースカフェ
8　※料理系サイト（クックパッド，オレンジページ，楽天レシピなど）
9　ななつ星／サン・プリンセス
10　個人間スキル取引
11　“ブランド香水”ミニ柔軟剤
12　※スマート～（スマートシティ，スマートハウス，スマート家電など）
13　ネオ裏原系
14　新・希少糖ドリンク
15　シナプソロジー
16　※次世代テレビ（4K 対応テレビ，次世代テレビ，有機 EL テレビ等）
17　「抗ロコモ」ギア＆フード
18　スマホ風機能追加デジカメ
19　伊勢・出雲詣で
20　デジタル対応上質ノート

※ ALBERT 抽出商品

4.　おわりに

本研究では，Twitter 解析に際し，実際にツイートデータを収集する方法や，解析に利用できる技術を紹介し，それらを用いて 2013 年のヒット予測を試みた．Twitter などのトレンド分析においては，特定の話題が何らかの原因によって急激に増加するというバースト現象が起きるが，その要因についても具体的に分析し，時系列のトレンドを加味して，最終順位の予測を行った．今回は 1 年の 1/3 である 1 月～4 月までの 4 カ月間の Twitter 解析であり，これをもって年間の予測を行うことはきわめて乱暴なことであり，またどのようなワードを抽出するかによってツイート総数が大きく影響を受けてしまうため，年末の発表時に再度分析を試み，予測精度を上げていくことが今後の課題である．

参考文献

[1]「2013 年ヒット予測ランキング」，『日経トレンディ』，日経 BP 社，2012.
[2] “クチコミ@係長,” http://www.hottolink.co.jp/kakaricho.（2013 年 5 月 6 日アクセス）
[3] “REST API Rate Limiting in v1.1,” https://dev.

twiter.com/docs/rate-limiting/1.1.（2013 年 5 月 6 日アクセス）
[4] “REST API v1.1 Resources,” https://dev.twitter.com/docs/api/1.1.（2013 年 5 月 6 日アクセス）
[5] “otterapi”, http://code.google.com/p/otterapi/wiki/Resources.（2013 年 5 月 6 日アクセス）
[6] “Using the Twitter Search API,” https://dev.twitter.com/docs/using-search.（2013 年 5 月 6 日アクセス）
[7] A. V. Aho and M. J. Corasick “Efficient String Matching: An Aid to Bibliographic Search,” *Communications of the ACM*, **18**(6), 333–340, June 1975.
[8] J. Kleinberg, “Bursty and Hierarchical Structure in Streams,” In Proc. 8th SIGKDD, 91–101, 2002.

