Vol.2010-DBS-150 No.2
Vol.2010-IFAT-99 No.2
2010/8/4

Twitterにおける流行語先取り発言者の検出システムの開発

白木原 渉[†1] 大石哲也[†2] 長谷川隆三[†3]
藤田 博[†3] 越村三幸[†3]

情報検索エンジンでは最新の情報，特に流行している事柄を検索するのは難しい．近年，Twitterが急激に普及し始めた．Twitterでは，世の中で流行している事柄(流行語)について，多くの人が発言する傾向がある．Twitterのユーザーの中でも特に流行に敏感な人(trendspotter)を知ることができれば，その人の発言に注目することで，流行している事柄についての情報をさらに簡単に手に入れることができる．

本システムを実現する手法として，一般のバースト検出アルゴリズムを用いたが，これがTwitterの発言に対しても利用できることがわかった．さらに，本システムによって，5277人のユーザーの中から，24人のtrendspotterを抽出することに成功した．

TRENDSPOTTER DETECTION SYSTEM FOR TWITTER

Wataru Shirakihara,[†1] Tetsuya Oishi,[†2]
Ryuzo Hasegawa,[†3] Hiroshi Hujita[†3]
and Miyuki Koshimura [†3]

It is too difficult for us to find out trends with search engines. Twitter, a popular microblogging tool, has seen a lot of growth since it launched in October, 2006. Information about the trends are posted by many twitterers. If we find out trendspotters from twitterers, and follow them, we can get it more easily.

Our system uses the burst detection algorithm, and we verified its effectiveness for Twitter's posts. Finaly, we succeeded in detecting the 24 trendspotters by 5277 users.

1. はじめに

ここ十数年の技術の向上により，インターネットは，情報検索のツールとして大きく発展してきた．欲しい情報があるとき，多くの人がインターネットの情報検索エンジンに頼っている．しかし，情報検索エンジンで検索される情報は，誰かが更新するまでは，もとの情報のままであり，最新の情報，特に流行している事柄を検索するのは難しい．

近年，SNSの一種であるTwitterが急激に普及し始めた．日本におけるTwitter利用者数は，2010年1月現在473万人(2010年2月，ネットレイティングス調べ)である[3)]．

ただし，これらの数字は，ウェブサイト（http://twitter.com）から利用している人たちだけを集計したものにすぎず，携帯電話からの利用者や専用クライアントからのユーザーは含まれていない．したがって，実際のTwitter利用者数は，この数字よりもはるかに多い．

Twitterは他のSNSと異なる特徴を多く持っており，Twitterを「情報交換のためのツール」として利用しているユーザーが多い．よって，Twitterをうまく利用することで，流行している事柄をより簡単に知ることができる．

次節で述べるが，Twitterは，流行している事柄を知るためのツールになり得る特徴を多く持っている．そのため，世の中で流行している事柄（流行語）について，多くの人が発言する傾向がある．Twitterには多くのユーザーがいるが，その中でも特に流行に敏感な人を知ることができれば，その人をフォローすることで，流行している事柄についての情報をさらに簡単に手に入れることができる．

第2節でTwitterについて述べ，第3節で関連研究について述べる．第4節で本システムの具体的な動きを述べ，第5節では実際にシステムを動かして得た結果を示し，その結果に対する考察を行う．さらに，得られた流行に敏感な人たちがその後も流行に対して敏感であるかを確かめる評価実験とその結果・考察についても述べる．第6節で結論を述べ，第7節で今後の課題とともに本論文をまとめる．

†1 九州大学大学院システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
†2 九州大学情報基盤研究開発センター
Research Institute for Information Technology, Kyushu University
†3 九州大学大学院システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University

Vol.2010-DBS-150 No.2
Vol.2010-IFAT-99 No.2
2010/8/4

2. Twitter

Twitter についてより詳しく述べる．

2.1 Twitter

Twitter(ツイッター) とは，個々のユーザーが 140 文字以内の「つぶやき」（以下発言とする）を投稿する，いわば「ミニブログ」である．他のユーザーを登録（フォロー）すると，その人が書き込んだ発言を読むことができる．また，自分がフォローされることもある．フォローした人の発言は，書き込まれた順番に「タイムライン」と呼ばれる，自分の専用画面に表示される．さらに，誰が誰をフォローしているかが一覧でき，相手にメッセージを送ってコミュニケーションをとることができる．

2.2 Twitter の特徴

Twitter の特徴として，以下の５つが挙げられる．

(1) RT（ReTweet）の伝播力
RT とは，ReTweet（リツイート）の略語であり，他人の発言を自分の発言にそのまま引用して紹介する際，引用を表す目印のようなものである．RT が繰り返される間には，コメントが付け加えられたり，元情報の間違いが正されるなど，フィルタリングの効果ももたらしている．結果，情報の信頼度とスピードが高められることになった．

(2) 短縮 URL
Twitter はしばしば，ニュースサイトやブログのアドレスを掲載するために使われる．しかし，Twitter ではそれらもまた 140 字の発言の中にカウントされてしまうので，URL をそのまま掲載していては，他のことが何も書けなくなる．
それを解決したのが，URL を短縮するサービスである．URL を含む発言が 140 字を超える場合，URL が自動で短縮されるようになっている．元の長い URL に転送するためのアドレスを生成し，短い字数で提供している．

(3) ボット（bot）
ボット (bot) と呼ばれる，プログラムを使った自動書き込みが存在する．定時に決まったことを発言したり，企業が自社ブログの更新情報を機械的に告知したりするために使われることが多い．例えば，JR の運行状況や，TV のこれから放映される番組を知らせてくれるボットがいる．これらのボットをフォローすることによっても，最新の情報を得ることが可能となった．

(4) API 開放によるカスタマイズ
2006 年に Twitter のサービスが始まると同時に，Twitter の API を利用して開発したクライアントツールが多く登場した．クライアントツールとは，Twitter にアクセスするための専用ソフトで，ウェブでアクセスするよりも使い勝手のよい操作性を提供してくれるものである．
また，Twitter の機能をうまく使ったウェブサービスも多く登場している．そのひとつに，buzztter[7] がある．buzztter は，Twitter で短期間に多くの人が発言した語を流行語とみなし，それを抽出してユーザーに提供するサービスである．
本研究では，この buzztter を利用してシステムを構築する．その詳細については後述する．

(5) ハッシュタグ
ハッシュタグとは，頭にハッシュマーク (#) が付けられたキーワードを，発言の中に付け加えることによって，Twitter 上で特定の話題を検索しやすくしたものである．例えば，Twitter の検索機能を使って「#traindeley」を検索すると,「#traindeley」を含む発言が一覧表示され，そこには電車の遅延の情報が並ぶ．
さらに現在では，#を付けて書き込まれた文字列はハイパーリンクとして表示される機能が付加されたため，このハッシュタグをクリックするだけで発言の一覧が表示される．

2.3 他のネット系メディアとの比較

本研究では Twitter を用いるが，他のネット系メディア (ブログ，他の SNS) と比較をしながら，Twitter を選んだ理由を述べる．

2.3.1 ブログとの比較

Twitter がブログと異なる最大の点は，基本的に自分のタイムラインを眺めているという点である．ブログのコンテンツはブラウザを経由して，ユーザーが自分から見に行かなければならない．したがって，Twitter の方がより気軽に利用できる．

また，情報発信ツールとしても，Twitter はブログと異なる．ブログではじっくり意見を書くことができるが，読む側は長い文章を読まなければならない場合が多い．一方 Twitter では，140 字という字数制限があるため，要点を簡潔にまとめられた発言を読むことができる．これも，Twitter の気軽さの一因である．

2.3.2 他の SNS との比較

身近な距離感であり，フォローによってユーザーを結ぶという点で，Twitter は SNS の

Vol.2010-DBS-150 No.2
Vol.2010-IFAT-99 No.2
2010/8/4

一つと考えることができる．しかし，mixi[8] など他の SNS ほど密接なコミュニケーションを求められているわけではない．基本的に承認なしでフォローでき，見知らぬ人の発言を RT できる．このような点で，Twitter は，SNS としてはかなりオープンなものだと言える．

また，他の SNS にはなく Twitter にあるものが，リアルタイム性である．他の SNS では，更新した記事をユーザーが自分から見に行かなければ，その記事を知ることはできない．一方 Twitter では，発言された瞬間にユーザーのタイムラインに表示されるため，他の SNS よりもかなり早い段階で情報を得ることができる．このような特徴から，Twitter ならば，数時間後のイベントの告知や目の前で起きていることを伝える手段となり得る．

以上の比較から，流行している事柄を知るためのツールとしては，Twitter は他のネット系メディアよりも優れていると言える．

3. 関 連 研 究

関連研究として，バースト検出について述べる[4]．

ブログや電子掲示板に対する書き込みなどでは，ある話題に注目が集まる場合に特定の語句の出現頻度が急激に上がるといった現象が起こる．これは，多くの人がその話題に注目し，言及することによって，その話題に関連する固有名詞などが一時的に出現しやすくなるためである．

Kleinberg の提案するバースト検出手法[5] は，ブログや電子掲示板への書き込みを，新聞記事などのように時間情報のついた文書の集合を意味する document stream としてとらえ，document stream 中で document 数が急激に増加している部分 (バースト) を発見する手法である．

我々は，このバースト検出手法が Twitter のデータ（発言）に対しても有効であることを確認し，これを用いてシステムを構築した．

4. 流行語先取り発言者の検出システム

我々が提案する流行語先取り発言者の検出システムの内部の動きを具体的に説明する．

(1) buzztter から流行語 s を得る．
buzztter には様々な流行語 (短時間の間に多くのユーザーが発言した語句) が抽出されており，そのなかには流行に関する語句とは違う語句，例えば「おはよう」「お昼」「そうです」などが存在するが，そのような語句は除外して抽出した．

(2) Twitter の API を用いて，流行語 s を含む発言を検索し，そのタイムライン (ユーザ名と発言した時刻) を得る．

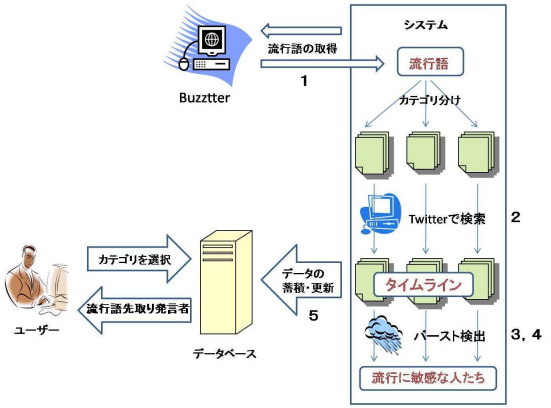

図 1 システムのイメージ
Fig. 1 Overview of the system

(3) 流行語 s を含む発言が急増 (バースト) する時間帯を 3 節のバースト検出アルゴリズムを用いてタイムラインから探す．

(4) バーストに含まれる発言のうち，最も早い発言 (番号 t が小さい発言)20 個を取り出し，その発言を行ったユーザー 20 名を流行語 s に対する trendspotter とする．

(5) 1〜4 を繰り返し，一定数以上の流行語に対して trendspotter となった人を検出する．

5. 実　　験

本研究で提案したシステムが，実際に trendspotter を検出できるかどうかを確かめる実験を次の 3 段階に分けて行った．

(1) 発言に対するバースト検出の有効性の確認 (5.1 節)
本システムにおける trendspotter の検出の手法として，バースト検出アルゴリズムを用いるが，これが本システムに適しているかを確認する実験を行った．

(2) trendspotter の検出 (5.2 節)
一定数以上の流行語に対して trendspotter となるユーザーが，実際に得られるかどうかを検証する実験を行った．

(3) 検出した trendspotter の有効性の確認 (5.3 節)

Vol.2010-DBS-150 No.2
Vol.2010-IFAT-99 No.2
2010/8/4

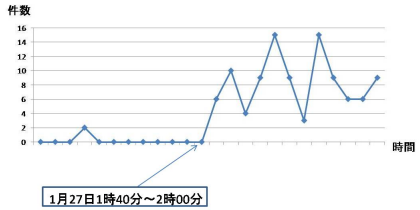

図 2 流行語「更生法」の発言数
Fig. 2 A number of posts including the buzz word "更生法"

5.2 節で得た trendspotter が，その後も流行に敏感であるかどうかを確かめる実験を行った．

次節より，各実験について結果と考察を交えて述べていく．

5.1 実験１：発言に対するバースト検出の有効性の確認

実際に，buzztter から得た流行語のうち，３つ (「更生法」「有楽町」「グラミー」) を例にとって，twitter の発言に対してバースト検出が有効であることを確かめた．

まず，それぞれの流行語に対して得られた結果を示し，その後考察を述べる．

5.1.1 結果１：「更生法」

2010 年 1 月 27 日早朝，株式会社 WILLCOM が会社更生法を申請したことを受け,「更生法」が流行語となった．流行語「更生法」を twitter で検索して得たタイムラインの，2010 年 1 月 26 日 22 時 00 分から 27 日 6 時 00 分までの発言数をグラフにすると図 4.1 のようになった．グラフの横軸は時間 (20 分区切り)，縦軸はその 20 分間でなされた発言数である．

このタイムラインをバースト検出アルゴリズムで解析した結果，1 月 27 日 2 時以降の発言がバースト状態にあるという結果が得られた．

5.1.2 結果２：「有楽町」

流行語「有楽町」を twitter で検索して得たタイムラインの，2010 年 1 月 26 日 7 時 00 分から 26 日 6 時 00 分までの発言数をグラフにすると図 4.2 のようになった．グラフの横軸は時間 (20 分区切り)，縦軸はその 20 分間でなされた発言数である．

この例では，１つの流行語「有楽町」で，1 月 26 日 9 時 00 分～10 時 00 分，1 月 26 日 18 時 00 分以降，の２つのバーストが検出された．

5.1.3 結果３：「グラミー」

流行語「グラミー」を twitter で検索して得たタイムラインの，2010 年 1 月 28 日 16 時 00 分から 2 月 1 日 10 時 00 分までの発言数をグラフにすると図 4.3 のようになった．グラフの横軸は時間 (20 分区切り)，縦軸はその 20 分間でなされた発言数である．

この例でも，結果２と同じように，１つの流行語「グラミー」で，1 月 31 日 21 時 20 分～2 月 1 日 1 時 40 分，2 月 1 日 7 時 00 分以降，の２つのバーストが検出された．

まず，結果に対する考察を各流行語に対して行い，最後に全体を通しての考察を述べる．

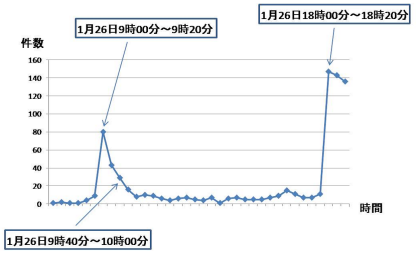

図 3 流行語「有楽町」の発言数
Fig. 3 A number of posts including the buzz word "有楽町"

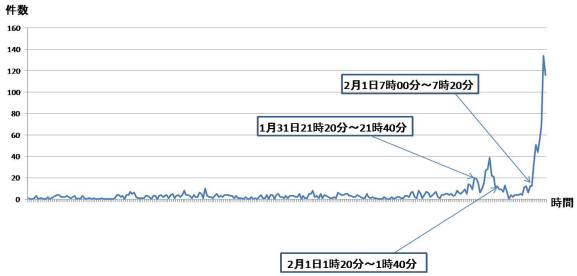

図 4 流行語「グラミー」の発言数
Fig. 4 A number of posts including the buzz word "グラミー"

5.1.4 考察１：「更生法」

流行語「更生法」の結果に対して考察を行う．

- 図 5 中の時刻 (a) においては，若干の発言数の増加が見られるが，バースト検出アルゴリズムでは，この増加はバーストと認識しなかった．
  流行語とは，多くのユーザーがある一定の期間の中で発言する語句であるので，時刻

Vol.2010-DBS-150 No.2
Vol.2010-IFAT-99 No.2
2010/8/4

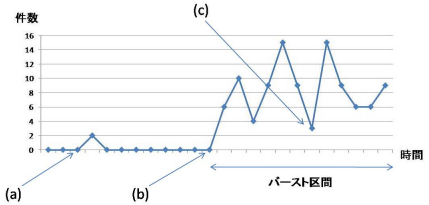

図 5 考察 1：流行語「更生法」の結果に対する考察
Fig. 5 Consideration1: ”更生法”

(a) のように，少数のユーザーが発言した語句については流行語と認識しないほうが適切である．

以下は，時刻 (a) での実際の発言である．

Tue, 26 Jan 2010 23:04:31 +0900	
mubot	http://twitter.com/mubot
福岡銀行の学界更生法の適用申請を受けて、タス通信争奪戦を繰り広げているビルマジャンパー磁気とチュニジアメールアドレス習わしの 2 社が 19 日、それぞれ支援表明を発表した。	

この発言は，先述した話題 (株式会社 WILLCOM の会社更生法申請) とは違う話題であることがわかる．

今回の例 (更生法) のように，流行となる話題と違う話題で発言され得る語句が流行語となる場合があるが，本システムでは流行に関する発言だけを抽出することが求められる．そのため，流行とは関連しない発言を除外できるという点で，バースト検出アルゴリズムは有効である．

- 時刻 (b) で急激な発言数の増加が見られるが，バースト検出アルゴリズムも，この地点からバーストが始まると認識した．
- 時刻 (c) においては，発言数が短い時間の間に大きく減少・増加しているが，バースト検出アルゴリズムは，(c) の前後のバーストは同一のものであると認識した．以下は，時刻 (c) の前後の発言である．

Wed, 27 Jan 2010 04:01:40 +0900	
hagexx	http://twitter.com/hagexx
まじですか？　【速報】ウィルコム、会社更生法申請へ	
Wed, 27 Jan 2010 04:06:41 +0900	
ysbee	http://twitter.com/ysbee
「ＰＨＳ最大手のウィルコムは…機構とソフトバンクの支援を前提に更生法の適用を申請するプレパッケージ＝事前調整型の法的整理手法を活用」RT @sarustar RT @Hagexx: まじですか？【速報】ウィルコム、会社更生法申請へ - http://ow.ly/10FS9	
Wed, 27 Jan 2010 04:08:47 +0900	
awazeno999	http://twitter.com/awazeno999
ニュースだと「ウィルコムが会社更生法申請」の台詞だけが独り歩きするだろうから、ますます客離れ進むかも。	
Wed, 27 Jan 2010 04:23:17 +0900	
takeori	http://twitter.com/takeori
ウィルコムばずってると思ったら会社更正法申請だった。:【速報】ウィルコム、会社更生法申請へ http://bit.ly/bW56yI	
Wed, 27 Jan 2010 04:24:12 +0900	
tdaiki	http://twitter.com/tdaiki
RT @takeori: ウィルコムばずってると思ったら会社更正法申請だった。:【速報】ウィルコム、会社更生法申請へ http://bit.ly/bW56yI	
Wed, 27 Jan 2010 04:26:13 +0900	
tabloid	http://twitter.com/tabloid
な、なんと !!　RT ウィルコムばずってると思ったら会社更正法申請だった。:【速報】ウィルコム、会社更生法申請へ http://bit.ly/bW56yI (via @takeori)	

04 時 08 分から 04 時 23 分の間に発言がないため，この 20 分間は発言数が急激に増減している．しかし，発言の内容を見ると，同じ話題に対する発言であることがわかる．1 つの流行語を含む発言数が，短時間の間に急激に増減することがあり得る．しかし，ある 1 つの流行語を含む話題が 2 つ存在し (例えば，１つ目が A 社の更生法申請の話題，２つ目が B 社の更生法申請の話題)，この短時間の間にそれが切り替わることは考えにくい．したがって，短時間の間の急激な増減の前後を，違うバーストとして扱うと不適切である．

Vol.2010-DBS-150 No.2
Vol.2010-IFAT-99 No.2
2010/8/4

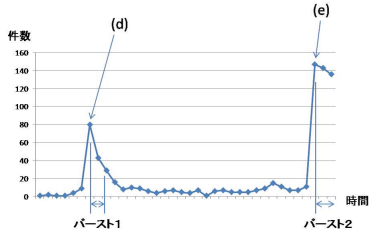

図 6 考察 2：流行語「有楽町」の結果に対する考察
Fig. 6 Consideration2: ”有楽町”

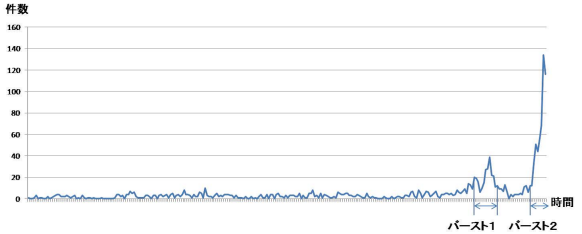

図 7 考察 3：流行語「グラミー」の結果に対する考察
Fig. 7 Consideration3: ”グラミー”

5.1.5 考察 2：「有楽町」

流行語「有楽町」では，2つのバーストが得られた (図 6)．バースト 1 中の時刻 (d) における発言と，バースト 2 中の時刻 (e) における発言を比較する．

- 時刻 (d)

Tue, 26 Jan 2010 09:10:44 +0900	
togamim	http://twitter.com/togamim
有楽町線ストップや。信号トラブル@新富町	

- 時刻 (e)

Tue, 26 Jan 2010 18:01:57 +0900	
MAKEPURA	http://twitter.com/MAKEPURA
「西武百貨店の有楽町店が閉鎖」が NHK ニュースのトップニュースだった。そんなに大きいの？	

時刻 (d) における発言は，東京地下鉄有楽町線の信号トラブルについての発言であり，また，時刻 (e) における発言は，西武有楽町店の閉店についての発言である．

このように，1つの流行語が 2 つの流行している話題のどちらにも含まれることがあるが，本システム場合，別々の「流行」としてとらえる必要がある．したがって，これら 2 つが別々のバーストに分けられることは，本システムにおいて適切である．

5.1.6 考察 3：「グラミー」

流行語「グラミー」におけるバーストを図 7 に示す．

この例では，2010 年 1 月 28 日 16 時 00 分から 2 月 1 日 10 時 00 分までという，比較的長い期間からバースト検出を行ったが，発言数が急激に増加している 2 つの時間帯を適切にバーストと認識した．

5.1.7 全体の考察

以上 3 つの例に対する考察から，バースト検出アルゴリズムは Twitter の発言に対しても用いることができ，本システムの実現において以下の点で有効であると言える．

- 急激に発言数が増加 (バースト) した時間帯は，バースト検出アルゴリズムによってもバーストと認識される．
- 微小な発言数の増加によってはバーストと認識しない．
- 短時間の間に急激な発言数の増減があった場合，その前後を違うバーストと認識することはない．
- 1つの流行語が 2 つの流行している話題のどちらにも含まれる場合でも，それぞれを異なるバーストとして認識する．
- 比較的長い期間のデータからも，適切にバーストを検出できる．

5.2 実験 2：trendspotter の検出

本システムでは，一定数以上の流行語に対して trendspotter となるユーザーを検出するが，実際にそのようなユーザーが存在するかどうかを確かめる実験を以下の要領で行った．

- Buzztter より流行語を 200 件抽出．
- 検索して得る発言数は最大 1500 件．
- 実験期間は 2010 年 1 月〜2 月
- 各流行語に対してバースト検出を行い，trendspotter を各バーストから 20 人検出する．
- 各ユーザーが何件の流行語に対して trendspotter であったかを調べる．

Vol.2010-DBS-150 No.2
Vol.2010-IFAT-99 No.2
2010/8/4

5.2.1 実験２の結果

流行語 200 件に対し，5277 人の trendspotter が検出できた．１つの流行語に２つ以上のバーストが検出される場合があったため，200 件 ×20 人 =4000 人 よりも多く検出された．N 件の流行語に対して trendspotter となったユーザーの人数と，その人数の全体 (5277 人) に対する割合を表 1 にまとめた．

表 1 N 件の流行語に対して trendspotter となったユーザー数と，全体に対する割合

Table 1 A number of trendspotters for N buzz words and percentage of total

該当件数 (N)	人数	割合 (%)
1	4734	89.71
2	463	8.77
3	56	1.06
4	12	0.23
5	9	0.17
6	3	0.06

5.2.2 実験２に対する考察

実験２で得られた結果に対する考察を行う．

表 1 に示したように，多くのユーザーの中から，流行に対して特に敏感な人が検出できた．例えば，３件以上の流行語に対して trendspotter となったユーザーは，全体の 1.52% であり，5277 人のユーザーの中から，80 人の特に流行に敏感な人を検出することができた．

以上の実験・考察から，本システムは，当初の目的「流行に敏感な人を検出する」を達成できたと言える．

5.3 実験３: 検出した trendspotter の有効性の確認

実験２で，trendspotter が実際に検出できることを示したが，そのユーザーたちが，その後も流行に敏感であるかどうかを確かめる実験を以下の要領で行った．

- Buzztter より流行語を 110 件抽出
- 検索して得る発言数は最大 1500 件
- 実験期間は 2010 年 2 月 5 日〜2 月 7 日 (実験２を行った後の期間)
- 各流行語に対してバースト検出を行い，バースト期間の間に発言したユーザーを抽出する．
- 各ユーザーが，何件の流行語に対してバースト期間中に発言したかを調べる．その際，全ユーザーの平均件数と，実験２で検出した trendspotter 上位 24 名の平均件数を比べる．

5.3.1 実験３の結果

流行語 110 件に対し，34381 人のユーザーがバースト期間内に発言した．まず，N 件の流行語に対してバースト期間内に発言したユーザーの人数と，その人数の全体 (34381 人) に対する割合を表 2 に示す．

表 2 N 件の流行語に対してバースト期間内に発言したユーザー数と，全体に対する割合

Table 2 A number of users posted N buzz words in bursts and percentage of total

該当件数 (N)	人数	割合 (%)
2 件以下	31174	90.67
3〜5	2916	8.48
6〜8	227	0.66
9〜11	46	0.13
12 件以上	18	0.05

次に，実験２で検出した trendspotter 上位 24 名の，N 件の流行語に対してバースト期間内に発言したユーザーの人数と，その人数の全体 (24 人) に対する割合を表 3 に示す．

表 3 trendspotter 上位 24 名の N 件の流行語に対してバースト期間内に発言したユーザー数と，全体に対する割合

Table 3 A number of the trendspotters posted N buzz words in bursts and percentage of total

該当件数 (N)	人数	割合 (%)
2 件以下	8	33.3
3〜5	8	33.3
6〜8	4	16.7
9〜11	3	12.5
12 件以上	1	4.2

5.3.2 実験３に対する考察

実験３で得られた結果に対する考察を行う．

表 2 と表 3 を比較すると，全ユーザーと比べて，実験２で検出した trendspotter の方が

Vol.2010-DBS-150 No.2
Vol.2010-IFAT-99 No.2
2010/8/4

流行語を多く発言していることがわかる．例えば，流行語を 100 件中 6 件以上発言した人の割合は，全ユーザーでは 1% にも満たないが，実験 2 で検出した trendspotter の中では 33.4% と，非常に高い割合となっている．

また，バースト期間中に流行語を発言した全ユーザーは，1 人あたり平均 1.43 件の流行語を発言しているが，実験 2 で検出した trendspotter は 1 人あたり平均 4.54 件の流行語を発言していた．

以上の実験・考察から，本システムで得られた流行に敏感な人たちは，その後も，多くの流行語に対して，その語を含む発言をする傾向にあることがわかった．

6. 結　　論

本研究では，流行している事柄についての情報をさらに簡単に手に入れることを可能とするために，Twitter ユーザーの中から流行に敏感な人 (trendspotter) を検出するシステムを提案した．

このシステムを実現する手法として，バースト検出アルゴリズム (ブログなどの記事数の急増する時間帯を検出するアルゴリズム) を用いたが，これが Twitter の発言に対しても利用できることがわかった．さらに，急激に発言数が増加 (バースト) した時間帯は，バースト検出アルゴリズムによっても正しくバーストと認識されることもわかった．したがって，バースト検出アルゴリズムは本システムにおいて有効であり，推薦すべき trendspotter を適切に検出することが可能であると言える．また，本システムで得られた特に流行に敏感な人が，他のユーザーと比べて，その後も多くの流行語を発言することがわかり，本システムの有効性を示すことができた．

7. 今後の課題

- **バースト検出アルゴリズムの改良**
  本研究では，各流行語に対してバースト検出を行ったが，バーストを検出できない流行語が存在した．その原因として，バースト検出アルゴリズム中のパラメータの数値が適切でなかったことが考えられる．流行語を Twitter で検索して得られたタイムラインの性質 (総発言数や時間帯，期間など) によって，自動的にパラメータを調節するようなアルゴリズムを考案することによって，この問題を解決できることが予想される．
- **流行語のカテゴリ分け**
  本システムでは，検出した各 trendspotter が，どの分野の流行に敏感なのか (例えば，音楽や映画など) をその人が発言した流行語から判断し，カテゴリ分けを行う．Twitter のユーザーを分野別 (政治・IT・芸能スポーツなど) にクラスタリングし，さらにその中でも有用かつリアルタイム性の高い情報 (流行) を多く提供しているユーザーを抽出する．クラスタリングを行うためには，クラスタとクラスタを分ける基準が必要である．この基準を見つけるために，Twitter のデータ（発言/つぶやき）のマイニングを行う．具体的には，1）フォロー・被フォロー関係，2）発言に対する返信・被返信，3）ハッシュタグ，4）発言が行われた時間，5）発言間の関連度 (文章同士の近さ)，6）バーストの検出 (短時間内に多くの発言に含まれた語句の解析) などを考慮し，これらを組み合わせてクラスタリングし，さらに目的のユーザーを抽出する．
  これらを実行する際，膨大な計算時間を要することが予想されるため，分散処理のフレームワークである Hadoop を用いる．また，これらの情報を格納するデータベースとして，Hadoop と相性の良い HBase, Cassandra を用いる．

以上のような項目を今後の課題とし，システムの改善を目指す．

謝辞　本研究は科研費（21500102）の助成を受けたものである．

参 考 文 献

1) 神田 敏晶, “ Twitter 革命 ”ソフトバンク，2009
2) コグレ マサト，いしたに まさき, “ ツイッター　140 文字が世界を変える ”毎日コミュニケーションズ，2009
3) http://www.netratings.co.jp/New _ news/News02242010.htm
4) 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 ”， IPSJ SIG Notes，2004(23) pp.85-92 20040304
5) Jon Kleinberg, “ Bursty and hierarchical structure in streams ”the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2002
6) 長尾 真（編）, “ 岩波講座ソフトウェア科学 15 自然言語処理 ”，pp.568-576，岩波書店，1996
7) http://buzztter.com
8) http://mixi.jp