Twitterにおけるバーストの検出と生起要因に関する分析

筑波大学
図書館情報メディア研究科

2014年3月

水沼友宏

目次

表目次

図目次

1　研究の背景と目的

1.1　はじめに

情報技術の発達、スマートフォンの普及などにより、ソーシャルメディアを通じた情報の収集・発信、コミュニケーションが一般的なものとなっている。その影響力も大きく、2010 年末から 2011 年にかけ、北アフリカ、中東諸国で起こった民主化運動、通称「アラブの春」では Twitter[*1]や Facebook[*2]などのソーシャルメディアが大きな役割を果たしたことが知られている。また、2008 年のアメリカ大統領選で当選したオバマ大統領は、選挙期間中、ソーシャルメディアを頻繁に活用したことが知られている [1]。日本でも 2013 年参議院選挙でのネット選挙解禁を受け、多くの政党や政治家がソーシャルメディアを活用した [2]。ICT 総研の調査 [3] によれば、2012 年末時点での日本のソーシャルメディアの利用者は 4965 万人に上り、ネット利用人口の半数以上がいずれかのソーシャルメディアを利用していることが報告されている。

ソーシャルメディアの一種である Twitter は、2006 年の登場以降ユーザ数を伸ばし、2012 年時点で、そのユーザ数は 5 億人を突破した [4]。また、Twitter 社の公式発表 [5] によると、2012 年 12 月時点での月間アクティブユーザ数が 2 億人を超え、1 日に 4 億のツイートが投稿されている。日本においても、2008 年の日本語版発表を契機としてユーザ数が増加し、さらに 2011 年 3 月の東日本大震災以降、非常時における有効な情報収集・コミュニケーションツールとしても認識されるようになった。2012 年 2 月時点で、日本国内の Twitter アカウント数は、2990 万件であり、この数はアメリカ、ブラジルについで第三位である。また、アカウントごとに発信されるツイート数では、日本は、オランダに次いで第二位である [4]。

Twitter の、速報性、情報拡散性、簡便性といった特徴により、あるイベントが生起した際にツイート数が平常時と比較し大きく増加することがある。例えば、2013 年 8 月 2 日、テレビで「天空の城ラピュタ」が放送され、主人公が「バルス」と叫ぶと同時に多くのユーザがツイートを行っている。Twitter 社の公式発表によると、その際の 1 秒間におけるツイートの投稿数 (TPS: Tweets per Second) は、143,119 ツイートに上った。本論文では、このように、あるイベントが生起した際にツイート数が大きく増加する現象を、「バースト」と定義し、これについて定量的な分析を行い、Twitter 上でユーザが種々の社会現象をどのように捉え、伝達しているのか、また他の情報メディアとどのような関係にあるのかを明らかにする。具体的な目的は、(1) どのようにバーストが生起しているか否かを判断すべきか、(2) どうしてバーストが生起するのか、(3) バースト時の Tiwtter の投稿特徴はどのようなものか、 (4) それぞれのバーストをどのように分類できるのか、を明らかにすることである。

[*1] https://twitter.com/
[*2] https://www.facebook.com/

バーストは Twitter の特徴である「速報性」、「簡便性」、「双方向性」を象徴する現象であり、バーストの実態を明らかにすることは、Twitter のメディア特性をより鮮明なものとする。加えて、Twitter とは、他のソーシャルメディアに比べ、行動の自由度が高いメディアであるのにも関わらず、バースト時は多くの人々が一斉に情報を発信するという特異な状態であり、バースト現象の実態を明らかにすることは、人々の情報行動に関する一知見を与えることができると考えられる。

1.2 本論文の構成

本論文は、5 章から構成されている。まず、1 章で背景と目的、および Twitter についての概要を述べた後、2 章で既往研究について詳述する。3 章では、データ収集の方法及び、バースト閾値算出方法について説明する。さらに、4 章では結果と考察を述べる。4 章 1 節では、バースト時の投稿特徴を、バースト時と非バースト時を比較することによって明らかにする。2 節では、1 章で明らかにしたバースト時の投稿特徴をもとに、各バーストの分類を行う。さらに 3 節では、地震によって生起したバーストに影響を与える要因をロジスティック回帰分析によって明らかにする。4 節では、感情語をもとにバーストと感情の関係を明らかにする。最後に、5 章でまとめと今後の課題を述べる。

1.3 Twitter とは

Twitter は、ソーシャルメディア、および、マイクロブログの一種とされている。本節では、ソーシャルメディアや、マイクロブログの定義などから、Twitter を概説する。

『インターネット白書』によれば、“ソーシャルメディアとは、友人や知人らとのコミュニケーションや交流を促進する場あるいは仕組みで、友達やフォロワーといったつながりを有するインターネット上のサービスであり、具体的には、SNS やマイクロブログ”[6] と定義されている。また、Merrian-Webster には、“ソーシャルネットワークやマイクロブログのような電子的なコミュニケーションの一形態で、ユーザは、情報やアイディア、個人的なメッセージ、その他のコンテンツ (動画など) を共有するためのオンラインコミュニティを作成する (forms of electronic communication (as Web sites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (as videos))”[7] と定義されている。IT 用語辞典には、“ソーシャルメディアは、インターネット上で展開される情報メディアのあり方で、個人による情報発信や個人間のコミュニケーション、人の結びつきを利用した情報流通などといった社会的な要素を含んだメディアのこと。利用者の発信した情報や利用者間のつながりによってコンテンツを作り出す要素を持った Web サイトやネットサービスなどを総称する用語”[8] と記載されている。このように、定義はさまざまであるが、概して、ユーザ間のつながりやコミュニケーションを促進するインターネット上の情報サービスといえる。主なソーシャルメディアとそのユーザ数を表 1 に纏めた。ユーザ数はすべて 2013 年時点のも

のである。なお、ユーザ数の MAU とは、月間アクティブユーザ数 (Monthly Active Users) を示す。また、IR 資料とは、株主や投資家に対する (Investor Relations) 資料を意味し、IPO 申請書とは、新規株式公開 (Initial Public Offering) 申請書を意味する。

表 1 主なソーシャルメディア

名称	運営会社	ユーザ数	参照元
Facebook	フェイスブック株式会社	11 億 5000 万 (MAU)	IR 資料 [9]
Twitter	ツイッター社	5 億 2 億 1500 万 (MAU)	IPO 申請書 [10]
Google+	Google 社	3 億 (MAU)	Wall Street Journal on web [11]
LINE[*3]	LINE 株式会社 (旧 NHN Japan)	3 億	プレスリリース [12]
モバゲータウン[*4]	株式会社ディー・エヌ・エー (DeNA)	5,380 万	IR 資料 [13]
GREE[*5]	グリー株式会社	3,890 万	IR 資料 [14]
Ameba[*6]	株式会社サイバーエージェント	3,000 万	プレスリリース [15]
mixi[*7]	株式会社ミクシィ	1,250 万 (グループアプリ会員数)	IR 資料 [16]

Twitter は、SNS (ソーシャル・ネットワーキング・サービス) の一種とみなされることもある。SNS とは、広義には、社会的ネットワークの構築の出来るサービスやウェブサイト、とされる。例えば、谷口は、“SNS (Social Network Service) とは社会的ネットワークをインターネット上で構築するサービス”[17] と述べている。Twitter 上には、情報収集・拡散の目的のためだけではなく、実社会や、ウェブ上の友人とつながり、コミュニケーションを行っているユーザも少なくない。それゆえ、社会的なネットワークが構築されていると解釈でき、この意味では Twitter は SNS に含まれる。事実、谷口 [17] は SNS の例として、mixi や Facebook とともに、Twitter を挙げている。その他にも、Twitter を SNS の一部と認識している調査・研究は少なくない [18][19][20]。一方、狭義には、ソーシャル・ネットワーキング・サービスとは人と人とのつながりを促進・サポートする、コミュニティ型の会員制のサービス、と定義されている。Twitter は、双方向ネットワークだけでなく、一方向のネットワークも存在することからコミュニティ型のサービスであるとは限ら

ず、また会員制のサービスとも言い難いため、狭義の意味には含まれない。IT用語辞典は、SNSを、“人と人とのつながりを促進・サポートする、コミュニティ型のWebサイト。友人・知人間のコミュニケーションを円滑にする手段や場を提供したり、趣味や嗜好、居住地域、出身校、あるいは「友人の友人」といったつながりを通じて新たな人間関係を構築する場を提供する、会員制のサービスのこと。”[21]というように、狭義の意味で定義している。それゆえ、併記されている例も、Facebook、LinkedIN、GREE、mixiにとどまり、Twitterは含まれていない。また、Twitter社の技術担当副社長のMichael Abbott[22]は、Twitterはソーシャルではなく、情報ネットワークであると述べ、SNSとしてのTwitterを否定している。このようにSNSの定義は曖昧であり、TwitterがSNSに含まれるか否かについても様々な考えが存在するが、本研究では狭義のSNSの定義を採用し、TwitterはSNSに含めない。

ソーシャルメディアの一種とされる「マイクロブログ」は、インターネット上で、不特定多数または特定の人に向けてごく短いメッセージを発信したり、他の人のメッセージを読んだりすることができるサービスの総称である。マイクロブログ以外にも、「ミニブログ」や「簡易ブログ」という呼称も存在するが、本研究では、マイクロブログに統一する。Twitterは、投稿あたりの文字制限が140文字と一般的なブログと比べると、制限文字数が極めて短く、代表的なマイクロブログサービスの一つとされている。Twitter以外のマイクロブログとしては、中国で主に使われているweibo[*8]などがある。また、mixiのmixiボイスや、Amebaの Amebaなうなど、SNSの一部のサービスとして提供される場合もある。ブログに比べ、投稿できる文字数は少ないが、多くのモバイル端末で使用できることや、短い投稿が主であることから、外出中にも比較的簡単に投稿できる。また、短いテキストで構成されるため、更新が容易であり、リアルタイムなコミュニケーションが可能である。

Twitter社はアメリカに本社を持つ会社である。2006年3月に、Jack Dorsey、Evan Williams、Biz Stone、Noah Glassにより2006年3月に開発され、同年7月にウェブサイトが公開された。2007年3月にアメリカで開催されたイベント「SXSW: サウス・バイ・サウスウェスト」のブログ部門で賞を受賞したことにより、1日あたりの投稿数が20,000件から60,000件に急増した[23]。2008年には、3ヶ月あたりの投稿数が1億件を突破し[24]、同年には、インターフェースが日本語かつ日本語での投稿が可能な日本語版もリリースされた。2010年には、BlackBerry端末、iPhone端末、Android端末用の公式クライアントアプリ「Twitter」の無料配布を開始し、さらに、2013年には、ニューヨーク証券取引所に株式を上場した。

Twitterは、140文字以内の投稿を入力し、他のユーザと共有する無料のウェブサービスである。一般に、ユーザの投稿のことを、「ツイート」、または、「つぶやき」と呼び、各ユーザのホーム画面は、「タイムライン」と呼ばる。ツイートを投稿する際や、自分のタイムラインを閲覧する際には、パソコンや携帯電話などの端末から自身のアカウントにログインする必要がある。他のユーザを

[*8] weibo.com/

「フォロー」 (登録) することによって、他のユーザのツイートが自身のタイムラインに新着順に表示される。先に述べたように、マイクロブログはその文字制限もあいまって、他のユーザが投稿したツイートが、そのユーザをフォローしているユーザのタイムラインに即座に現れる「速報性」を持つ。Twitter についても例外ではなく、Twitter の特性として、即時性、リアルタイム性が挙げられる。さらに、マイクロブログの特性として、専門的な知識を持たずとも、アカウント取得や投稿が容易であること、また、携帯電話やタブレット端末など利用端末の多様さによる「簡便性」が挙げられる。マイクロブログの中でも特に、Twitter は、様々な端末に対応した無料のクライアントアプリが Twitter 社から配布されており、より簡便性に優れていると言える。また、Twitter は他のユーザをフォローするために、ほとんどの場合で承認がいらないことや、つながりに相互性が必要ないことなどから、他のソーシャルメディアに比べ、ユーザ同士が気軽につながることができる。これに加え、他のユーザのツイートを自分のフォロワーのタイムラインに表示させる「リツイート」などの機能により、情報が広範囲に拡散される「情報拡散性」を持つ。

1.4 Twitter に関連する用語の解説

この節では、本論文中に言及される Twitter の機能や用語について簡単に説明を行う。

1.4.1 ユーザ名とアカウント名

ユーザー名は、@で始まる英数字のもので、アカウントのプロフィールページの URL に表示されるアカウント特有のものである。ログイン、リプライ、およびダイレクトメッセージで使用される [25]。ユーザ名は半角英数字または「_」(アンダーバー) で構成されなければならないというシステム上の制限が存在する。一方、アカウント名は、ユーザ名とは別にプロフィール内に表示される ID である。ユーザ名とは異なり、アカウント名は漢字やひらがな、記号、絵文字なども使用でき、他のユーザのアカウント名と同一でもかまわないなど、設定の際の自由度が高い。

1.4.2 非公開アカウント

非公開アカウントとは、ツイートを非公開にしているアカウントである。鍵付きアカウントと呼ばれる場合もある。図 1 に示すように、非公開アカウントの場合、ユーザ名の隣に鍵のアイコンが表示され、ツイート内容は表示されない。非公開アカウントをフォローするには、ユーザから承認を受ける必要があり、承認を受けたユーザのみがツイートを閲覧することができる。非公開アカウントを持つユーザが作成したツイートは、パブリックタイムラインに表示されない。また、フォローの承認を受けているユーザであっても公式リツイートを行うことはできない。

図 1　鍵付きアカウント

1.4.3　リツイート (Retweet, RT)

リツイートとは他の誰かのツイートを再投稿することを指す。RT と略して表記される場合もある。他のツイートをリツイートすることで、そのツイートが自分のフォロワーのタイムラインにも表示されるため、情報の拡散を促す機能であると言える。

リツイートは公式リツイートと非公式リツイートの 2 種類に大別できる。リツイートが公式機能ではなかった 2009 年以前は、慣習的に、引用するツイートの前に「RT @ユーザ名」を付することで、他人のツイートの引用であることが示されていた。なお、ここに示される@ユーザ名は、リツイートされたユーザ (以下、被引用ユーザとする) のユーザ名であり、これによって元ツイートの発信者を特定できる。しかし、この方法では、引用元のツイートが投稿された時間が分からない、被引用ユーザがツイートの削除を試みても、引用されたツイートまでは削除できないといった問題を抱えていた。これらの問題を解決するため、2009 年 11 月に英語版の一部ユーザーを対象にリツイートの公式機能が試験的に実装され、2010 年に日本語版に実装された。公式リツイートは、各ツイートに付されているリツイートボタンを押すことにより、利用することができる (図 2)。公式リツイート機能が用いられたツイートには、リツイートを行ったユーザのアイコンではなく、リツイートされたユーザのアイコンが表示される。また、リツイート元のツイートが削除されると、すべての公式リツイートも削除される。このように、Twitter の公式リツイートボタンを用いたリツイートを公式リツイートと呼び、それ以外の方法で引用がなされる場合は、非公式リツイートと呼

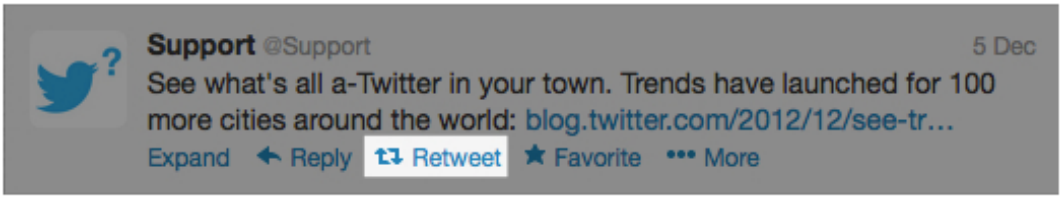

図 2　リツイート機能

ばれる。

公式リツイート機能が実装された後も、非公式リツイートは用いられている。例えば、非公開アカウントを引用したい場合、公式リツイート機能を使用することができないため、ツイートのコピーに、「RT」表記を付与することで、引用がなされている。なお一般にこの場合は、非公開アカウントのユーザに配慮し、ユーザ名は付与されない。また、リツイートに自分のコメントを付与したい場合は、図 3 に示すように、非引用ツイートをコピーし、その前に「RT @ユーザ名」を付与し、さらにその前にコメントを付与する。なお、この方法は、Twitter 公式のヘルプセンター [25] にも書かれており、Twitter 社はこの方法を公認していると言える。また、非公式リツイートを行う場合は、「RT」表記以外に、引用 (Quote) の略式表記である「QT」が用いられることもあるが、一般に、QT を用いるのは、コメント付きの非公式リツイートを行う場合に限られる。

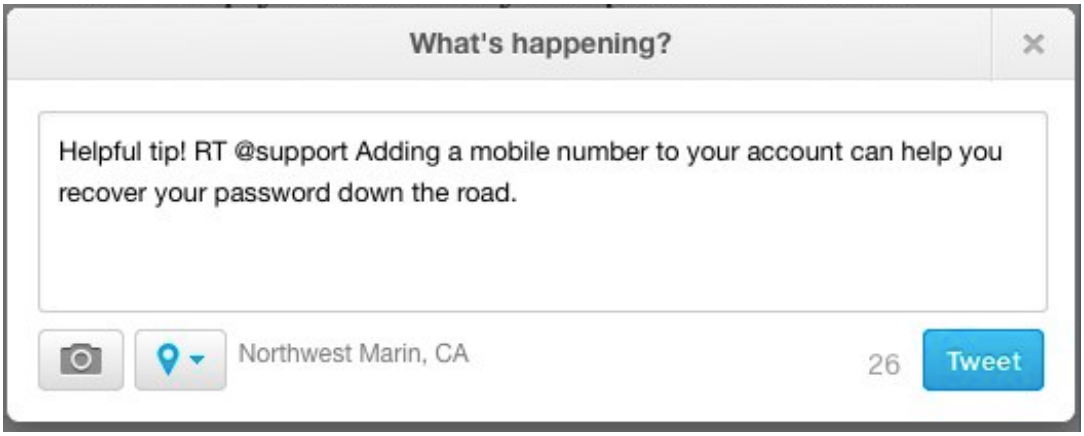

図 3　非公式リツイート

1.4.4　**リプライ** (Reply, @)

特定のユーザー名 (@..) から始まるツイートが「リプライ」という [26]。リプライという呼称以外にも単に「@ (アット)」と呼ばれる場合もあり、特定ユーザ名を付与してツイートすることは「リプライを送る」や「@を飛ばす」と表現される。リプライは、特定のユーザに対するツイート

であり、リプライ先ユーザのタイムライン及び、リプライ元ユーザとリプライ先ユーザの両方をフォローしているユーザのタイムラインにしか表示されない。このため、リプライ機能は公開範囲を狭める仕組みであるとも言える。ユーザ名が文頭ではなく文中に表記される場合は、メンション (Mention) と呼ばれ、リプライとは区別される。メンションは、通常のツイートと同様に、ツイートを投稿したユーザをフォローしているユーザのタイムラインにツイートが表示される。

1.4.5　**ハッシュタグ** (hashtag, #)

ハッシュタグとは#記号と、半角英数字で構成される文字列のことである [26]。発言内に「#(テキスト)」を含め投稿すると、その記号つきの発言が検索画面などで一覧できるようになる。図 4 では、ツイート末尾の「#高尾山」、「#紅葉」がハッシュタグである。URL 及びハッシュタグの色が他のツイート本文とは異なることからも分かるように、Twitter 社の公式機能となった後は、ハッシュタグにリンクが貼られるようになり、ハッシュタグをクリックすると、そのハッシュタグが含まれたツイートの検索結果が表示される。なお、2011 年にハッシュタグに日本語を用いることが可能になった。また、携帯電話からは、ハッシュタグの検索結果ページから、そのハッシュタグを含めてツイートすることが可能である。なお、リツイート、リプライ、ハッシュタグはすべて、ユーザ間で自然発生的に使用されるようになったルールを、Twitter が公式機能として拡張したものである。

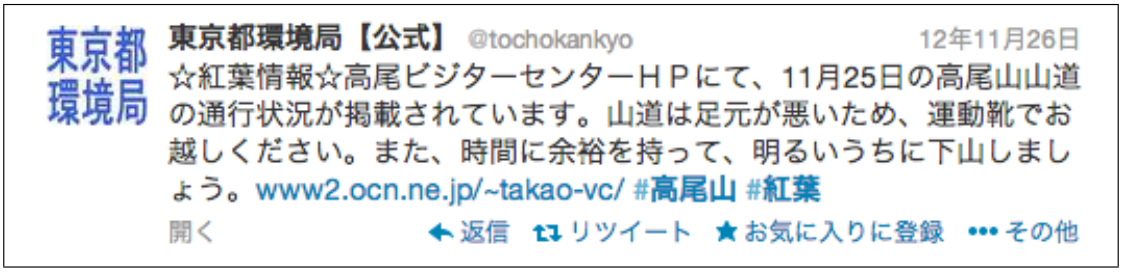

図 4　ハッシュタグ

2 既往研究

Twitter が普及するにつれて、Twitter がどのようなツールであり、人々がどのように Twitter を使用しているかを明らかにする研究は多く行われるようになっている。その嚆矢として、Java[27] は、2007 年 4 月 1 日から 5 月 30 日までの 2 ヶ月分のツイートデータを収集し、分析を行った。Java はツイート数 1,348,543 ツイート、ユニークユーザ数 76,177 ユーザのデータを分析し、投稿動機や Twitter のネットワーク構造を明らかにした。結果、(1) フォロー・フォロワーネットワークの直径は 6 であること、(2) ユニークユーザのうち 37,183 ユーザ (48.8%) がプロフィールに位置情報を記載しており、うち 56.6% が北米地域のアカウントであること、(3) 最も多いツイート内容は、「今、自分が何をしているか」であること、 (4) 全ツイートの 21% が@の付与されたリプライツイートであり、13% が URL を含んだツイートであること、 (5) リンク構造の解析から、ネットワークのハブであり多数のフォロワーを持つ情報源アカウントや多数のアカウントをフォローするものの、自らはほとんどツイートしない情報収集アカウントの存在などが識別されたことを明らかにしている。Krishnamurthy ら [28] はフォロー情報をもとにデータの収集を行うアルゴリズムと、ツイートをもとにデータの収集を行うアルゴリズムの 2 つのアルゴリズムを用いてツイートデータ及びプロフィール情報の収集を行っている。収集したデータを用いて Java らと同様の分析を行い、収集アルゴリズムの異なるデータセット間で、結果の比較を行っている。

特に、ユーザ間の関係や情報伝播という点に着目した研究は少なくない。Kwak ら [29] は、2009 年 6 月 6 日から 6 月 31 日にかけて、14 億 7000 万のフォロー・フォロワー情報と 4170 万のプロフィール情報を収集し、77.9% が一方向のフォローで、相互フォローは 22.1% に過ぎなかったことを報告している。さらに、67.6% のアカウントがいずれのユーザからもフォローされていなかった。以上のことから、Twitter はコミュニケーションツールというよりもむしろ情報源としての性質が強いことが示唆された。また、PageRank に基づいたユーザのランキングはフォロワー数と同様の結果を示していた一方で、リツイートのランキングでは、PageRank やフォロワー数とは異なるランキングが示された。また、Wu[30] は、リストを用いてユーザを、有名人 (celebrities)、ブロガー、ニュースメディア、組織などに分類し、カテゴリごとにフォロー・フォロワーネットワークやリツイート、リプライネットワークの分析・比較を行っている。分析の結果、有名人は有名人と相互フォローが行われるなど、同じカテゴリーのユーザは相互フォロー率が高いこと、ブロガーは、他のカテゴリに比べリツイートを行いやすいことなどが明らかとなった。Poblete[31] が、236 カ国の 4,736,629 ユーザから投稿された 5,270,609,213 ツイートを収集し、投稿特徴やネットワークの国際的な比較を行っている。日本のユーザについては、ユーザあたりのツイート数が多いものの、リツイート (RT) 、リプライ (@) 、ハッシュタグの使用比率が総じて低いこと、相互フォロー比率は 32.0% と分析対象国の中で最も高いことなどが示された。

ユーザ間の関係のなかでも特に、フォロー・フォロワーネットワークを対象とした研究として

は、Funda[32] は 2009 年 6 月と 2010 年 4 月の 2 時点でフォロー・フォロワーネットワークを収集し、個人間のユーザ間でのフォロー解消に関わる要因を示している。分析の結果、相互フォローの場合はフォロー解消がされにくいことや、フォロワーの多いユーザはフォロー解消されやすいことなどが示された。

Twitter 空間の情報伝播、特にリツイートに着目した研究としては、上述の研究の他に、768,000 ツイートの分析によりリツイート行動をモデル化した Sofus[33] の研究や、リツイートに着目し、非常時のユーザの行動を分析した Mendoza[34] の研究などがある。また、Suh[35] は、7400 万のツイートを分析し、リツイートされやすいツイートの特徴を明らかにしている。分析の結果、ハッシュタグが用いられていること、URL が付与されていることなどがリツイートに強い影響を与えることを示した。リツイート以外の方法で、情報伝播の実態を明らかにした研究は、Ye[36] の研究が挙げられる。彼は、Twitter 上の情報の伝播を Michael Jackson 死亡のニュースの広がりの分析によって明らかにしている。加えて、Galuba[37] は URL に着目し、その URL を含むツイートの伝播を分析している。

その他、Twitter 上のユーザの行動や投稿の特徴などを示した研究として、Palo[38] は Twitter 上の質問を抽出し、分析を行っている。分析の結果、エンターテイメントに関する質問が最も多いことや、返答を受けた質問は全体の 20% に満たなかったことを報告している。また、Antoniades[39] は、短縮 URL の分析を、Abel[40] はプロフィールの分析を行っている。

以上のように、Twitter のデータを用いて、Twitter がどのように使われているかを明らかにする研究は様々存在する。しかしながら、ツイート数のバーストに着目する研究は極めて少ない。Twitter とは、「つぶやく」という言葉に代表されるように、人々が好きな時間に好きなことをつぶやくメディアである。140 文字という制限や、ほとんどの場合、被フォロワーの承認がなくともフォローが可能な点を鑑みても、Twitter は他のソーシャルメディアよりも行動の自由度が高いと言える。しかしながら、バースト現象が起こっている間、多くの人々が一斉に投稿している。このように、自由な環境であるにもかかわらず、人々が一斉に投稿する現象について、その実態を把握し、分析することは、Twitter の特徴をより鮮明なものとするとともに、人々の社会的行動に関する知見を与えられると考えられる。また、Twitter で人々が一斉にツイートする現象というのは、Twitter 上で人々が活発に情報を発信しているという状態であり、人々の情報行動を明らかにする上でも有益である。

Twitter におけるバーストに関する研究として、ツイート本文に出現する単語からバーストを検出する研究は、幾らか行われている。例えば、Diao ら [41] は、シンガポールのユーザを中心に収集した 2011 年 9 月 1 日から 11 月 30 日までのツイート 3,967,927 ツイートに対し LDA（latent Dirichlet allocation)、および、二つの LDA 改良アルゴリズム（UserLDA, TimeLDA）を用いて、バーストする単語を抽出し、抽出した単語からトピックの自動検出実験を行っている。比較の結果、改良アルゴリズムは、より精緻にユニークなトピックを検出することができたことが報告され

ている。また、白木原ら [42] は buzztter[*9] から流行語を取得し、Kleinberg の提案するアルゴリズム [43] を用いて、流行語を含むツイートが急増する時間帯を検出している。しかし、これらは、単語に着目し、その単語を含むツイートが増加している状態をバーストと定義しており、全体のツイート数が増加する状態について分析は行われていない。

全体のツイート数の増減に着目した調査・研究はほとんど行われていない。NEC ビッグローブ [44] は、Twitter の 1 日あたりのツイート数やツイートに含まれる単語をランキング形式で発表している。調査の結果、2013 年 12 月に最もツイート数が多かったのは大晦日であること、そして大晦日は「NHK 紅白歌合戦」などテレビ番組に関する話題に関連するツイートが増加しており、最も話題となった単語は「進撃の巨人」の主題歌に含まれる「イェーガー」であったことを報告している。Twitter 社は、1 秒間のツイート数である TPS (Tweets Par Second) を測定し、公式ブログにてそのランキングを公表している。しかし、これらは全体のツイート数や特定の単語が含まれているツイートの数によるランキングを示すにとどまり、より詳細な分析は行われていない。乾ら [45] は、3 月 11 日の東日本大震災発生時前後のツイート 179,286,297 ツイートを対象とした分析のなかで、ツイート数の増減に着目している。分析の結果、震災後 1 週間で 1 分当たりのツイート数が最も多かったのは、3 月 15 日の静岡県東部の最大震度 6 強の地震の時であり、次に多かったのは同日の三陸沖の地震発生直後であったことを報告している。このように、Twitter においてツイート数の増加に着目した分析は極めて少なく、期間や対象が限定されていると言える。本研究では 2011 年 12 月から 2013 年 1 月という 1 年以上の期間を対象に、機械的な手法で網羅的にバーストの検出を行う。さらに、検出されたバーストについて、ツイート数の増減に着目するだけではなく、平常時との比較や、クラスタリングやロジスティック回帰などの手法を用いて多角的な分析を行う。

[*9] http://buzztter.com

3 方法

3.1 データの収集方法

バーストの分析を行うにあたって、まず、ツイートデータを収集し、そのデータから各日時のツイート数を調べ、バーストが起こっているか否かを判断する必要がある。従って、本節では、ツイートデータの収集方法について述べる。

分析に使用するツイートは，Twitter の Search API [*10]を用いて収集した。日本語で記述されたツイートを収集するため、言語に「ja」（日本語）と、日本全域をカバーする位置情報[*11]とを検索条件として指定した。

ツイートに付与される位置情報には、ユーザのプロフィールに自由記述する「location」情報と投稿時に GPS 等の値を自動的に付与する「geocode」情報の 2 種類がある。位置情報を検索条件とすることで geocode が指定した範囲内にあるツイートが収集できる。geocode が付与されていないツイートは、location に記入された情報が参照される。

収集の条件を東京駅を中心とする半径 100km 圏内と指定し、ツイートの収集を行った予備実験では、geocode が付与されておらず、さらに location 情報として「あっち」や「このへん」、「地方」、「ひみつ」といった曖昧な内容が書かれている場合でも、データが収集されていた。このことから Search API では、location が実際の地名等と一致せず、日本語でツイートされている場合、location あるいは、geocode は東京とみなされると考えられる。このため、geocode が付与されていない場合や location に実際の地名を書いていないユーザのツイートでも Search API に上述の検索条件を指定することで収集が可能である。本研究では各々のツイートごとにツイート ID、投稿時刻、ツイート本文の内容等の情報を収集した。ユーザのフォローに関する情報、お気に入り、およびツイートを非公開に設定しているユーザのツイートは、収集にかかる制限が大きいことから、分析に用いるデータとして網羅的な収集は行っていない。

データの収集は、2011 年 11 月 16 日から 2013 年 2 月 15 日までの 14 ヶ月間とし、5,285,607,227 ツイートを収集した。ユニークユーザ数は、10,918,410 ユーザであった。収集したデータの基本統計は表 2 に一覧する。リツイート（RT）はツイート本文の先頭に「RT」という文字が見られる場合とする。これには、公式リツイート及び、ツイートの先頭に RT という文字列を付与したツイートが含まれる。コメントを付加した非公式リツイートや QT によるリツイートは含まれない。同様に、@で始まる場合のみリプライとし、文頭以外の文中にユーザ名が含まれる、いわゆるメンションは含まれない。文字数の平均値は 45.65 文字であるが、最頻値は 21 文字であった。

[*10] http://search.twitter.com/search.json
[*11] 兵庫県西脇市を中心とする半径 2,000km 圏内

表 2 収集データの基本統計

データ項目	値
ツイート数	5,285,607,227
ユニークユーザ数	10,918,410
文字数 (平均)	45.76 文字
(最頻値)	21 文字
リツイート比率	8.82%
リプライ比率	39.02%

3.2 外れ値の定義

バーストの分析を行うにあたり、まず、バーストの検出方法を定めなければならない。先に述べたように、バーストとは、ツイート数が平常時と比較し、大きく増加している状態のことである。このため、バーストは、ツイート数が正の方向に大きく外れている場合、つまり、正の外れ値 (異常値) といえる。

外れ値 (異常値) の定義は様々である。例えば、Grabbs は、外れ値 (異常値) を表す outlying 及び outlier について、“外れ値 (異常値) とは、そのデータが含まれるサンプル集合の他のデータから著しく逸脱しているもの (An outlying, or outlier, is one that appears to deviate markedly from other members of the sampele in which it occurs)”[46] と定義しており、単純に他の値から異常に外れている値を、外れ値 (異常値) としている。これに対し、 Hawkins は “異なるメカニズムによって生起した、他の観測値から大幅に逸脱した観測値 (An observation that deviate so much from other observations as to arouse suspicion that it was generated by different mechanizm) ”[47] というように、異なるメカニズムによって、他の値から外れた値を外れ値 (異常値) としている。さらに、Barnet らは “他のデータセットと矛盾したある観測値 (または観測値のなかの部分集合) (An observation (or subset of observations) which appears to be inconsistent with the reminder of the set of data) ”[48] と述べており、他のデータ集合と矛盾した、調和性の無いデータのこととしている。また、外れ値と異常値は、英語ではどちらも outlier であるが、既出のように、日本語に翻訳する際、「外れ値」及び「異常値」と 2 通りの訳が存在する。外れ値と異常値の区別については、単に他から外れている値のことを「外れ値」、原因が存在するものは「異常値」というように定義づけられることもあるが、これらは明確に区別できない場合もあるため、区別を行わず使われることも少なくない。そこで、本研究でも、外れ値と異常値の区別を行わず、以下、便宜上、外れ値という用語を用いる。

3.3 適応可能な手法の選択

外れ値の検出手法として、様々な方法が提言されている。主な外れ値検出手法とその概略を表 3 に示した。各手法について、以下に詳述し、さらに適応可能性について述べる。

表 3 外れ値の検出手法

手法名	概略	適応可否
教師有り学習	各データに外れ値か否かの正解を与えた集合をもとに、正解が未知のデータについて正解を付与する	×
教師無し学習	正解の無いデータ集合を何らかの基準を設けてそれを最適にするような出力が示される	×
頻度指定	外れ値となる頻度を指定し、それを下回る出現数のものを外れ値とする	×
パーセンタイル	順に並べて上から何 % というように、外れ値の範囲を決定する	×
ROKU	シャノンのエントロピーと AIC に基づき外れ値を求める	○
統計分布に基づく棄却検定	統計的分布に基づき、有意水準に基づく棄却領域に含まれる値を外れ値とする	○
MAD 法	MAD (絶対中位偏差, median absolute deviation) を用いて外れ値を検出する	○

まず、機械学習の手法の一つである教師有り学習は、一定のデータに対し、それが外れ値か否かの正解を与え、その正解集合をもとに、正解が未知のデータに対し正解を付与する。しかし、本研究では外れ値であるか否かの正解集合を定めるすべが無いため、適応が困難である。同じく機械学習の手法の一つである、教師無し学習とは、正解の無いデータ集合から、何らかの基準が設けられ、それを最適にするような出力が示される。例えば、教師無し学習の手法の一つであるクラスタリングは、入力データを、類似度が高いデータグループに分類する手法である。教師無し学習では、外れ値か否かという基準で結果が出力されるとは限らず、かつ、パラメータを客観的に決めなければならないため、適応は困難である。頻度指定は、外れ値となる値の出現頻度を指定し、それを下回る出現数のものを外れ値とする方法である。例えば、外れ値を、出現頻度が 1 のものと指定した場合、値の大小にかかわらず、データ集合のなかで 1 つしか含まれていない値が外れ値となる。しかし、この手法では、本研究で検出したい「ツイート数が通常時と比較して大きく増加している状態」以外も外れ値として検出される可能性が高く、不適当であると考えた。また、パーセンタイルは、データを数字の大小の順に並べて上位 α % というように、α を任意に設定し、この条件に含まれ

る値を外れ値とする手法である。パーセンタイルは、一定期間において常に一定程度の外れ値が存在するという仮定にもとづいているが、本研究では、そのような外れ値はバーストとは呼ばない。このため、パーセンタイルも本研究では用いることはできない。ROKU とは、シャノンのエントロピーと AIC に基づき外れ値を求める手法である。また、統計分布に基づく棄却検定とは、統計的分布に基づき、有意水準に基づく棄却領域に含まれる値を外れ値とする手法である。MAD 法とは MAD (絶対中位偏差: median absolute deviation) を用いて外れ値を検出する手法である。これら 3 手法は、バーストの検出に適応可能であると考えられる。なお、これらの手法については、3.4 で詳述する。そこで、本研究ではバースト検出にあたって、教師有り学習、教師無し学習、頻度指定、パーセンタイルを除外し、ROKU、統計分布に基づく棄却検定、MAD 法のいずれかを用いる。

ここで、統計分布に基づく棄却検定については、複数の手法が存在する。主な検定手法を表 4 に示した。3 σ 法、Smilnov-Grubbs 検定、Smirnov-Grubbs 検定を複数外れ値で適応できるように拡張した増山の検定手法、Cochran 検定、Dixon 検定、マハラノビス距離の 6 つの手法について、複数外れ値に適応できるか否か、投入データの形式、及び数式を示している。なお、マハラノビス距離については、1 次元の場合の数式を記載している。数式における各記号と各手法については、次の段落から順に詳述する。

表 4　統計分布に基づく棄却検定

	複数外れ値	投入データ	数式
3 σ 法	○	1 次元	$X_n > \overline{N} + 3\sigma$
Smirnov-Grubbs 検定	△	1 次元	$X_n > \overline{X} + (S_x * t_{(n-1,\alpha)})$
増山の検定	○	1 次元	$X_n > \overline{X} + (S_x * \sqrt{1+1/n} * t_{(n-1,\alpha)})$
Cochran 検定	○	行列	$T = \frac{k(k-1)\sum_{j=1}^{k}(G_j - \overline{G})2}{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2}$
Dixon 検定	×	1 次元	$Q = \frac{gap}{range}$
マハラノビス距離	△	多次元	$X_n > \overline{N} + \theta\sigma$

3 σ 法とは、あるデータ X_n について、次式を満たす場合外れ値とする手法である。なお、平均値を $\overline{N}$、標準偏差を σ とする。

$$X_n > \overline{N} + 3\sigma \tag{1}$$

Smirnov-Grubbs 検定と増山の検定は、どちらも標本平均と不偏分散を用いる手法である。Smirnov-Grubbs 検定は、ある値 X_n が 2 式を満たす場合、増山の検定は、X_n が 3 式を満たす場合に X_n が外れ値として検出される。なお、平均値を $\overline{X}$、不偏分散を S_x とし、$t_{(n-1,\alpha)}$ は、データ数 n、有意水準 α のときの t 検定値である。増山の検定は Smirnov-Grubbs を複数外れ値に適応できるように拡張したものである。小林は、“Smirnov-Grubbs の棄却検定は，外れ値が 1 個のときは，検出力が高いが，外れ値が 2 個以上存在する場合は，一方が他方を隠す (Masking effect) ことがあり検出力が低下する”[49] と指摘しており、この Masking effect を考慮し、複数外れ値に適応可能にしたものが、増山の検定である。バーストは、一定期間に複数回生起することも十分考えられるため、本研究では、Smirnov-Grubbs と増山の検定を比較した場合、増山の検定の方が妥当であると考え、増山の検定を選択する。

$$X_n > \overline{X} + (S_x * t_{(n-1,\alpha)}) \tag{2}$$

$$X_n > \overline{X} + (S_x * \sqrt{1 + 1/n} * t_{(n-1,\alpha)}) \tag{3}$$

Cochran の検定とは、表 5 のような n 組、処理数 k の行列データから、検定統計量 T を求め、それが、自由度 k のときの χ^2 検定値よりも大きければ外れ値と見なす。T は次式により求められる。ここで、$G_j = \sum_{i=1}^{n} R_{ij}$(j=1,2,⋯,k)、$L_i = \sum_{j=1}^{K} R_{ij} (i = 1, 2, \cdots, n)$ であり、G_j は処理 j の合計値、L_i は n 組の合計値である。

表 5　Cochran の検定投入データ例

	Treatment 1	Treatment 2	⋯	Treatment k
Block 1	X_{11}	X_{12}	⋯	X_{1k}
Block 2	X_{21}	X_{22}	⋯	X_{2k}
Block 3	X_{31}	X_{32}	⋯	X_{3k}
⋯	⋯	⋯	⋯	⋯
Block n	X_{n1}	X_{n2}	⋯	X_{nk}

$$T = \frac{k(k-1)\sum_{j=1}^{k}(G_j - \overline{G})2}{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2} \tag{4}$$

Cochran の検定は、投入するデータが 1 または 0 の 2 値でならなければならないため、バーストの検出手法として用いることができない。

Dixon の検定では、Q 値と Q 検定表 [50] を用いて外れ値を検出する方法である。Q 値は次式によって求まる。なお、gap とは、外れ値と思われる値と隣接する値の差の絶対値であり、range とは、データ集合のなかの最大値から最小値を引いた値である。

$$Q = \frac{gap}{range} \tag{5}$$

検出方法から分かるように、外れ値同士の距離が近ければ、Q 値が小さくなることが懸念される。バーストの検出においても、複数外れ値がある場合に、外れ値同士の距離が近いことは十分考えられるため、バーストの検出に Dixon の手法を用いることができない。関も “Grubbs 型、Dixon 型の検出は外れ値 1 個の場合の検出法としては検出力も高く、一定の評価を得ているが。その一方で検定統計量の型からも明らかなように、特に複数個の外れ値が $G_{(i)}$、$D_{(p,q,r,s)}$ を過小に評価させることが起こり得る”[51] と指摘している。

マハラノビス距離は、1 次元データの場合、マハラノビス距離を、2 次元データ以上ではマハラノビス平方距離を用い、その距離が一定の基準値を超えた場合に外れ値として検出する手法である。2 次元以上のデータの場合に用いられるマハラノビス平方距離は、標本平均と共分散から相関係数を考慮した距離を算出し、その距離が一定の距離 θ を超えた場合外れ値とする。一方、一次元の場合は、平方する必要がないため、マハラノビス距離が用いられるが、これは、ある値 X_n と平均との差を標準偏差で正規化した値のことを指し、その値がある距離 θ を超えると外れ値とする。これは、次式によって表すことができる。なお、$\overline{X}$ は X の平均値である。

$$\frac{X_n - \overline{X}}{\sigma} > \theta \tag{6}$$

6 式を見て分かるように、θ =3 とすると、3 σ 法と一致し、これら 2 手法を用いることは冗長であると考えられる。それゆえ、本研究では 1 次元データの外れ値として、一般的に用いられる 3 σ 法を選択し、マハラノビス距離は用いない。

以上のことから、統計値に基づく棄却検定による外れ値検出では、3 σ 法及び増山の検定の 2 手法を用いる。

最後に、既往研究において、バーストを検出する手法としては、Kleinberg のアルゴリズムがあげられる [43]。この場合のバーストとは、時系列データにおいて特定イベントに対する言及が急激に増加した場合を指し、Twitter におけるバースト研究は、多くがこの手法に依拠している [41][42]。しかし、この手法は、時系列データ及び、ある条件を持つデータの 2 つのデータが必要である。つまり、2 変数のデータが必要であるが、本研究ではツイート数のみからバーストを判断するため、適応できない。また、Kleinberg の手法は、特定の単語が含まれているツイートのみの増加を対象としている点や急激な増加のみを対象としているが、本研究におけるバーストはそのような状態とは異なる。本研究では特定の単語が含まれるツイートの増加ではなく、ツイート数が増加をバーストとし、また、前の時間と比べ増加している場合ではなく、平常時と比較し増加している場合をバーストとするため、Kleinberg の手法を用いることは不適当である。

以上のことから、バーストの検出には、ROKU、3 σ 法、増山の検定手法、MAD 法のいずれかを用いる。次節から、適応可能な各手法について詳述し、さらに実際にテストデータでバーストの検出を行い、バースト検出に最も適合する手法を選択する。

3.4 適応可能な各手法の説明

3.4.1 ROKU

ROKU[52] とは、主に生物学の分野で用いられる外れ値検出手法であり、情報エントロピーと AIC (Akaike's Information Criterion: 赤池の情報量基準) を利用して、マイクロアレイ・データに含まれる外れ値を検出する方法である。なお、マクロアレイ・データとは、DNA を基盤上に整列化させたものである。エントロピーとは、確率変数が持つ情報の量を表す尺度で、それゆえ情報量とも呼ばれる。x の時のエントロピー $H(x)$ は次式によって表される。なお、P_i は x=i となる確率である。

$$H(x) = -\sum_{i=1}^{n} P_i \log_2 (P_i) \tag{7}$$

ここで、他の値から大きく外れた値が存在する場合、エントロピーが低くなることに留意されたい。図 5 から図 9 は、Kadota ら [52] の論文に記載されている図であり、x 軸が各組織を、y 軸が発現レベルを表す。本研究で用いるデータでは、横軸に時間を、縦軸にツイート数を用いる。図 5 に示した図は、1 点のみが、高い値を示し、他の値が 0 だった場合を示す。この場合、その値の大きさに関わらず、P_i=1 であるため、$H(x) = \log_2 1 = 0$ となる。図 6 は 1 点のみが、高い値を示し、他は 0 ではないが、0 に近い低い値を示している。この場合、$H(x)$=1.45 とエントロピーも低い。図 7 は、$H(x)$ が高い場合である。左の図は、すべて同程度の値である場合、右の図は値にばらつきはあるものの、突出して高い値は存在しない場合である。このとき、$H(x)$=3.32 と図 5 や図 6 に比べ、エントロピー $H(x)$ は高い値であることが分かる。以上のことが、少数の値が高い値を示し、その他の値が 0 に近い値である場合、エントロピー $H(x)$ は低い値を示すことが分かる。それゆえ、エントロピーが低い場合を外れ値として検出することが有効である。

しかし、この方法では、外れ値ではない部分が 0 に近くない場合、つまり、ベースラインが高いなかで外れ値が存在する場合、エントロピーの値が高くなってしまい、外れ値が検出されづらいという課題がある。図 8 では、突出した値があるにもかかわらず、図 7 と同じ $H(x)$=3.32 を示しており、外れ値として検出されづらいことが分かる。本研究で用いるツイートデータも、バーストしている時間以外のツイート数が 0 になるということはほとんど無く、エントロピーのみを用いた外れ値検出ではバーストが検出されにくいと考えられる。そこで、Turkey の Biweight 推定法によって求めた近似直線との差をとることで、外れ値ではない値のベースラインを 0 に近づける。

ここで、Turkey の Biweight 推定法について簡単に説明する。Turkey の Biweight 推定法とは、

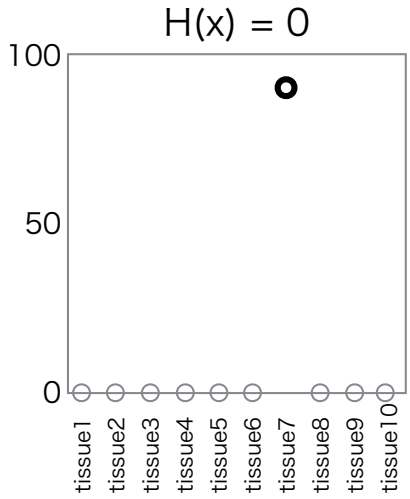

図 5 エントロピー $H(x)$ が 0 の場合

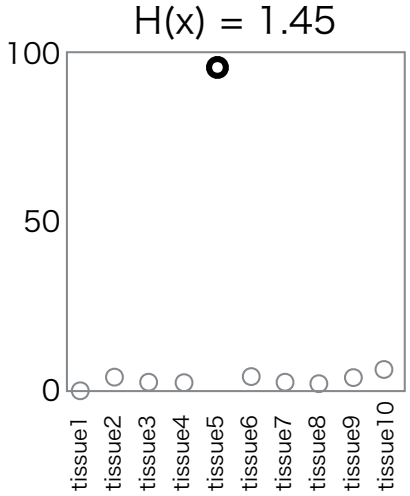

図 6 エントロピーが低い場合: $H(x)$=1.45

外れ値の影響を受けにくい近似直線を作る手法である。まず、各データを用いて、最小 2 乗法によって、近似直線を求める。次に、先に求めた近似直線からの距離を考慮して重み付けし、遠いほどウェイトを小さくする。さらに、重み付け済みデータで再び最小 2 乗法を用いて、近似直線を求める。このように、重みをつけたデータから最小 2 乗法を用いて外れ値の影響を受けにくい近似直線を求める方法を Turkey の Biweight 推定法という。最後に、元データからと近似直線との差をとることで、外れ値以外の値を 0 に近づける。これによって、エントロピー $H(x)$ は小さい値となる。図 9 において、左の図は、図 8 の値をもとに、Turkey の Biweight 法で求めた近似直線を示したものである。点線で示した線が近似直線である。さらに、右の図は、近似直線を引いたも

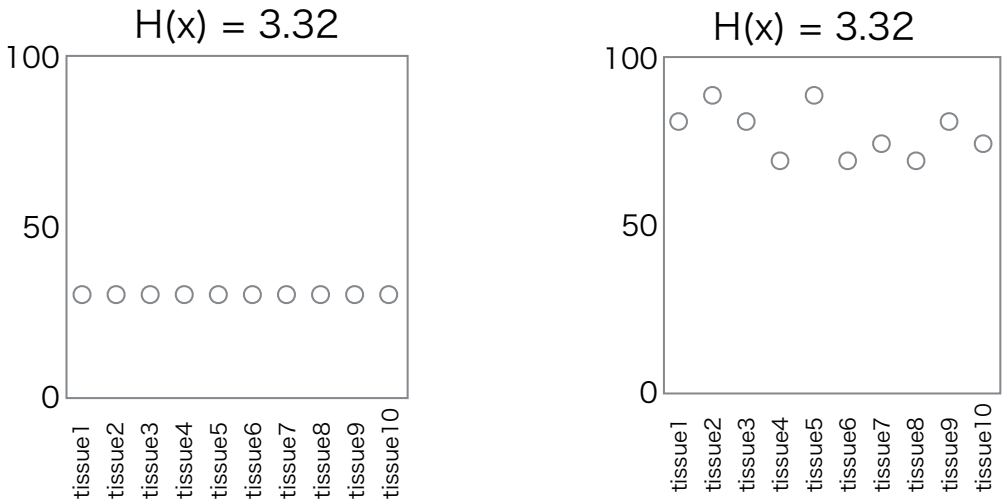

図 7 エントロピーが高い場合: $H(x)$=3.32

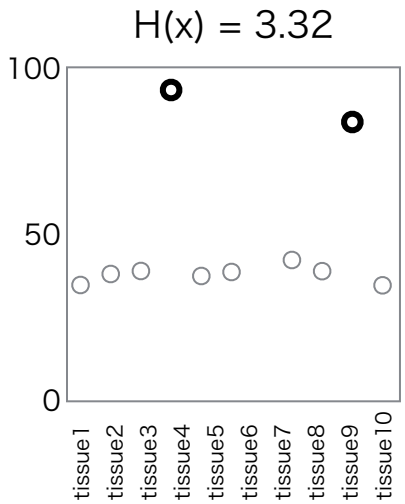

図 8 ベースラインが高い場合: $H(x)$=3.32

のである。図 9 を見て分かるように、左に示した Turkey の Biweight 法で求めた近似直線をを引く前はエントロピー $H(x)$=3.32 であるのに対し、右に示した近似直線を引いた後はエントロピー $H(x)$=1.74 とエントロピーが低いことが分かる。つまり、8 式に従って各データを変換することによって、ベースラインの高さの影響を受けずに外れ値の検出が可能となる。8 式において、T_{bw} は $x = i$ のときの Turkey の Biweight 法の値である。なお、負の値を取る場合はエントロピーを定義できないため、絶対値を取る。また、本研究とは直接関係はないが、絶対値を取ることで、不の方向への異常値も検出することができる。

$$f(x_{i)} = |x_i — T_{bw}| \tag{8}$$

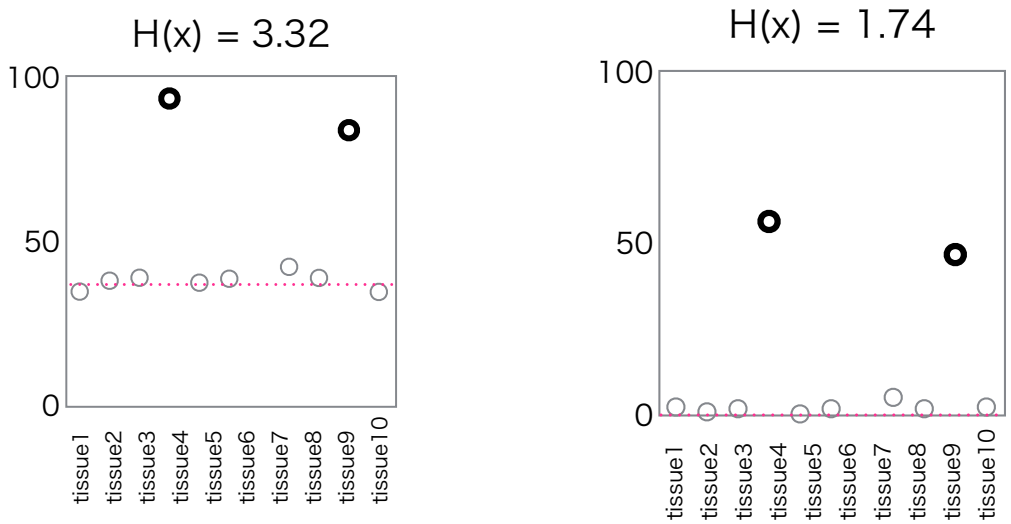

図 9　Turkey の推定法の適応

以上の方法で、エントロピーが小さいほど、外れ値が存在するというデータ集合を作成することが出来る。しかし、この方法では、外れ値の個数までは求められない。そこで、AIC を用いる。

AIC とは、モデルの良さを評価するための指標であり、次式によって定義される。ここで L は最大尤度、k は自由パラメータ数である。AIC 値が小さいモデルほど優良なモデルとされている。ここでの優良なモデルとは、精度が高く、ノイズの影響を受けすぎないようなモデルである。

$$AIC = -2InL + 2k \tag{9}$$

上位何個までを外れ値とするかという様々なモデルの AIC 値を求め、AIC 値が最も少ないものを選択する。その後、値の大きいものから、その個数だけ選択し、それが外れ値となる。

以上の方法が ROKU である。本研究では ROKU の分析には、R3.0.2 の ROKU 関数 [53] を使用した。なお、ROKU 関数を使用するにあたっては、発現変動解析用の R パッケージ TCC[*12]を読み込む必要がある。行列データを与え、ROKU で分析を行うと、「1: 特異的高発現、-1: 特異的低発現、0: その他」からなる「外れ値行列」を返す。特異的高発現とは、他の値に比べ高い値を示す外れ値である。一方、特異的低発現とは、他の値に比べ低い値を示す外れ値である。本研究ではバーストを通常よりもツイート数が増加している状態としているため、特異的高発現のみをバーストとする。バーストの検出を行うにあたっては、表 6 のように、行に時間当たりのツイート数、列に日にちを示した行列を作成する。バーストは 1 分ごとに検出を試みる。したがって、投入する行列の行数は 1,440 行 (60 (分) × 24 (時間)) であり、列数は日数と同数となる。このような形式の行列に ROKU を適応させると、表 7 のような結果が返される。先に述べたように、本研究では、特異的高発現のみを外れ値とするため、表 7 では、3 月 13 日の 23:58、23:59 を外れ値とする。

[*12] http://bioconductor.org/packages/release/bioc/html/TCC.html[54]

表 6 ROKU 投入データ例

	3 月 13 日	3 月 14 日	3 月 16 日	…	6 月 12 日
0:00	17770	16778	14725	…	19266
0:01	16067	15422	13127	…	17785
0:02	15819	15420	13209	…	17342
…	…	…	…	…	…
23:58	13648	14499	13886	…	15475
23:59	14359	14119	14151	…	16269

表 7 ROKU 結果例

	3 月 13 日	3 月 14 日	3 月 16 日	…	6 月 12 日
0:00	0	0	0	…	0
0:01	0	0	0	…	0
0:02	0	0	-1	…	0
…	…	…	…	…	…
23:58	1	0	-1	…	0
23:59	1	0	-1	…	0

3.4.2 3 σ法

3 σ法とは、あるデータ X_n が、以下の条件を満たすとき、X_n を外れ値とする手法である。

$$X_n > \overline{N} + 3\sigma \tag{10}$$

なお $\overline{N}$ はデータ集合 N における平均値、σ はデータ集合 N における標準偏差である。この手法は、工場の品質管理などで用いられる外れ値検出の手法であり、データが正規分布に従うと仮定した場合、平均値から 3 σを引いた値から平均値に 3 σを足した値の範囲に 99.7% の値が含まれ、この範囲に含まれない 0.3% 程度の出現率である値を外れ値としている。一般に 3 σ法では、X_n が平均値から標準偏差の 3 倍を引いたものより小さい場合も外れ値とするが、本研究ではツイート数が増加している場合のみ外れ値とするため、この場合は考慮しない。この手法を用いるにあたっては、まず、平均値に標準偏差の 3 倍を加えたものをバースト閾値として定め、そのバースト閾値に従ってバーストを検出する必要がある。例えば、表 8 のような 20 個のデータ集合が与えられた場合、平均は 13.2, 標準偏差は 9.75 であるため、平均に標準偏差の 3 倍を加えると 42.45 であり、この 42.45 という値がバースト閾値となる。このバースト閾値より大きい値、つまり表 8 のデータ

集合では 53 が外れ値として検出される。

表 8 投入データ

5	5	6	9	9	10	10	10	11	11
12	12	12	13	13	14	14	15	20	53

3.4.3 増山の検定手法

増山の検定手法 [55] は、$t_0 \geqq t_{(n-1,\alpha)}$ のとき、外れ値とする。なお $t_{(n-1,\alpha)}$ はデータ数 n、有意確率 α のときの統計量であり、統計量は t 分布に従う。t_0 は 11 式によって求める。11 式は、平均値を $\overline{X}$、不偏分散を S_x とする。不偏分散 S_x は 12 式により求められる。

$$t_0 = \frac{|X_n - \overline{X}|}{\sqrt{1 + 1/n} * S_x} \tag{11}$$

$${S_x}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{12}$$

11 式より、ある値 X_n が、次式よりも、大きければ外れ値とする。

$$\overline{X} + (S_x * \sqrt{1 + 1/n} * t_{(n-1,\alpha)}) \tag{13}$$

なお、本来の定義では次式よりも小さい場合も外れ値として認識されるが、今回はバーストの定義上、この場合は含めない。

$$\overline{X} - (S_x * \sqrt{1 + 1/n} * t_{(n-1,\alpha)}) \tag{14}$$

表 8 のようなデータ集合が与えられた場合、不偏分散は 10.00 であり、有意確率を 5% とすると、$t_{(19,0.05)}$=2.09 であるため、21.45 より大きい値、つまり表 8 のデータ集合では 53 が外れ値として検出される。

3.4.4 MAD 法

MAD 法 [56][57] とは、絶対中位偏差 (以下、MAD (median absolute deviation)) を用いる外れ値検出手法である。ある値 X_n が次式を満たす時、外れ値とする。

$$\frac{|X_n - med(X)|}{MAD} > 5 \tag{15}$$

なお、med(X) は、n 個のデータ、$X_i \cdots X_n$ の中位数である。MAD は med ($|X_i$-med$(X_i)|$) であり、それぞれのデータから中位数を引いた値の絶対値の集合の中位数である。もし、$X_i \cdots X_n$ が正規分布 (M,σ^2) からの無作為標本ならば、15 は $|X^* - med(X_i)| > 3.3725\sigma$ となり、中位数から偏差の絶対値が約 3.4σ 離れている値を外れ値と見なす。同様の手法で、3σ 離れている場合を外れ値とする方法も Maronna[58] により提唱されている。この場合、次式のように、15 式の右辺が 4.45 となる。

$$\frac{|X_n - med(X)|}{MAD} > 4.45 \tag{16}$$

本研究では、15 式を用いるが、ここで 3σ を考慮した方法が提唱されていることからも、3σ 法が、慣習的によく用いられる外れ値の手法であることが分かる。15 式より、ある値 X_n が次の条件を満たす場合、場合 X_n を外れ値とする。

$$X_n > 5MAD + med(X) \tag{17}$$

表 8 のようなデータ集合が与えられた場合、中位数は 11.5 であり、それぞれのデータから中位数を引いた値の絶対値のデータ集合が表 9 である。このデータ集合の中位数は 1.75 であり、20.25 よりも大きい値、つまり表 8 のデータ集合では、53 が外れ値として検出される。

表 9　各データから中位数を引いた値

6.5	6.5	5.5	2.5	2.5	1.5	1.5	1.5	0.5	0.5
0.5	0.5	0.5	1.5	1.5	2.5	2.5	3.5	8.5	41.5

3.5　各手法を用いたバーストの検出

どの手法を用いるかを決定するため、実際に、ROKU、3 σ 法、増山の検定、MAD 法を用いて、4 ヶ月分の平日ツイートデータを用いてバーストの検出を試みる。

先に述べたようにバーストの検出は分ごとに行うが、図 10 は、平日と休日の 1 日あたりの平均ツイート数の推移をグラフにしたものである。なお、休日とは、土曜、日曜、国民の祝日、お盆休み、正月休みとする。正月休みは、行政機関の休日に関する法律 (昭和六十三年十二月十三日法律第九十一号) 第一条に “次の各号に掲げる日は、行政機関の休日とし、行政機関の執務は、原則として行わないものとする。… (中略) … 三. 十二月二十九日から翌年の一月三日までの日” と定められているため、本研究でもこれに準拠し、12 月 29 日から 1 月 3 日までを休日とする。盆休みは、休日として法律では定められていないが、ツイート数の分布を見ると休日に近い分布を示してい

た。それゆえ、多くの人々が盆休みを取っており、ツイート数の分布も休日の分布に近い、8 月 13 日から 16 日をお盆休みとした。図 10 からも分かるように、平日は出勤前の 8 時前後と昼食時となる 12 時台にツイート数が増加するのに対し、休日は平日のような朝、昼のツイート数の増減が見られず、夕方にかけて徐々にツイート数が増加している。このように平日、休日のツイート数の違いを加味し、平日と休日では異なるデータセットを用いる。また、ROKU 以外の方法は、データ集合内で統一の閾値を設定し、その閾値をもとにバーストの検出を行う。しかし、図 10 を見て分かるように、時間によってツイート数の違いが見られる。例えば、朝 4 時台はツイート数が 2,000 件に満たない。一方、22 時、23 時は 14,000 件を超えている。このようなツイート数の違いを加味し、閾値に基づきバーストの検出を行う 3 σ 法、増山の検定、MAD 法は、分ごとに閾値を設定する。また、その時間における最大値、最小値は予め閾値算出のデータセットからは除外する。

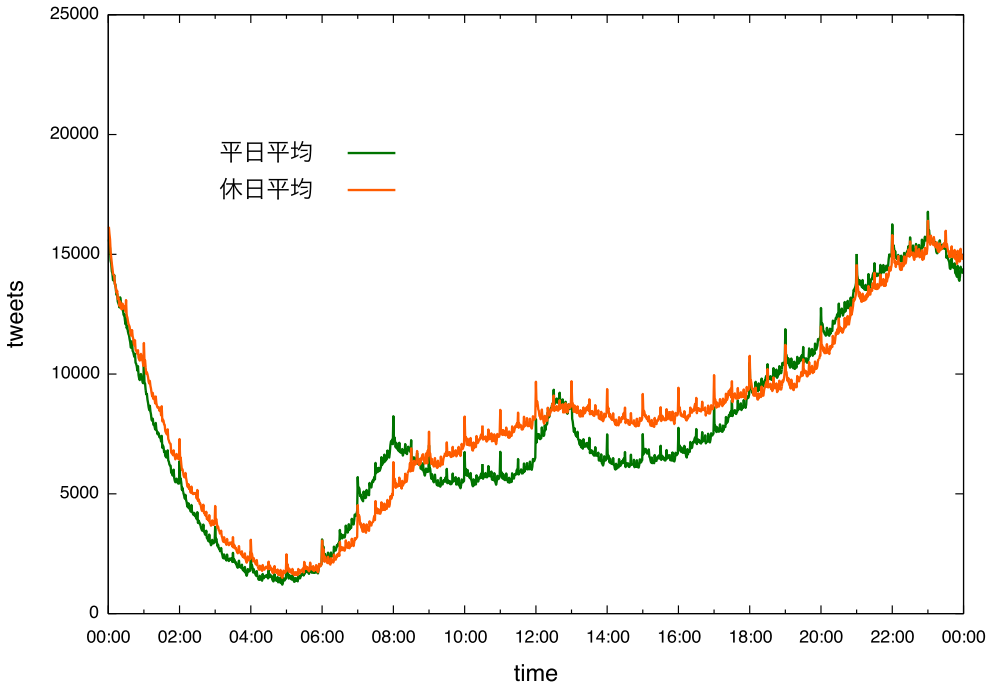

図 10　平日・休日のツイート数の推移

4 手法を用いたバースト検出に用いるテストデータ集合は、3 月 13 日から 6 月 12 日までの平日データ (67 日間分) とする。期間中に検出されたバースト数を表 10 に示した。表を見て分かるように、増山の検定では、有意確率 1% 及び 5% で検出を試みたが、1% では、他のデータと比較し検出数が大幅に少ないため、以後の分析では 5% を用いる。

表 10　各手法の検出数

手法名		検出数 (分)
ROKU		1,504
3 σ		1,557
増山の検定	(5%)	1,535
	(1%)	1,290
MAD 法		1,594

3.5.1　各手法間の関係

各手法間の一致率を、表 11 に示した。ROKU と MAD 法は一致率が 83.5% と最も高い。また、3 σ は、増山の検定、ROKU、MAD 法とすべてにおいて、80% 前後とある程度高い一致率を示す。増山の検定と ROKU の一致率は 71.8%、増山の検定と MAD 法の一致率は 70.0% 程度と、増山の検定は他の検定との一致率は低い。

表 11　2 手法間の一致率

	3 σ 法	増山の検定	MAD 法
ROKU	82.1%	71.8%	83.5%
3 σ 法		78.4%	78.4%
増山の検定			70.0%

各手法間の差集合の個数を表 12 に示した。なお、行に示す方法から列に示す方法を引いた差集合の個数を示しており、例えば、ROKU から 3 σ 法を引いた差集合の個数は 124 であったことを示す。ROKU から MAD 法を引いた差集合の個数が 94 と最も少ない。逆方向の MAD 法から ROKU を引いた差集合の個数も 184 と比較的少ない。また、3 σ と他の手法との差集合の個数は、MAD 法から 3 σ 法を引いた場合のみ、209 と 200 をわずかに超えるものの、それ以外の差集合はすべて 200 以内と比較的少ない。MAD 法と増山の検定、ROKU と増山の検定は、いずれも大きく上回る。

以上の結果から、ROKU と MAD 法は検出されたバーストが類似していること、3 σ は他のすべての手法と比較的類似した検出結果を示していることが分かる。

図 11 は他の 3 手法では検出されなかったが、その手法では検出されたバーストの個数、つまりその手法特有のバーストの個数を時刻別に示したものである。MAD 法は 6 時から 12 時以外の時間で、他の手法に比べ、バーストが検出されやすいことが分かる。また、3 σ 法は 14 時から 17 時までの時間でバーストが検出されやすいものの、その他の時間ではほとんど特有のバーストが見ら

表 12　行-列の差集合

	ROKU	3 σ法	増山の検定	MAD 法
ROKU		177	265	184
3 σ法	124		176	209
増山の検定	234	198		306
MAD	94	172	247	

れないことが分かる。

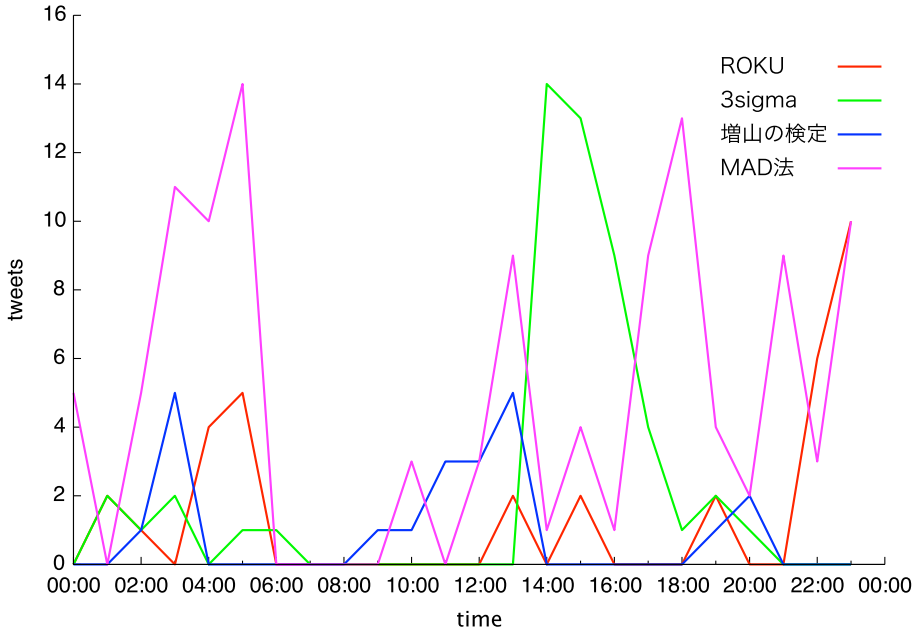

図 11　時刻別各手法特有のバーストの個数

3.5.2　各手法の比較

各手法の基本統計を表 13 に示した。3 手法の中で、平均継続時間が最も長いのは、3 σ法の 6.85 分である。また、最大値も 338 分と最も長い。平均継続時間が長いことは、長時間にわたる 1 つのイベントを、複数イベントとみなしづらいと考えられ、3 σ法はそのような傾向が見られることが示唆される。また、他の 3 手法の積集合とは、当該手法以外の、3 方法のバーストの積集合であり、これと当該手法を含めた 4 つの積集合との一致率が高いことは、その手法が他の手法の積集合を乱さないことを意味する。なお、4 手法の積集合の個数は 1,226 個であり、当該手法以外の積集合の個数で割ったものが一致率である。一致率が最も高い手法は 3 σ法の 99.2% である。ここで、検出個数が多ければ、この一致率は高くなりがちであるが、検出数は MAD 法が最も多いことに留意

されたい。以上のことから、4 手法の中で 3 σ法が最も理想的な手法であると考え、以下の分析では、3 σ法を用いる。

表 13　各手法の比較

	ROKU	3 σ法	増山の検定	MAD 法
検出数	1,504	1,557	1,535	1,594
連続バースト数	234	227	232	269
平均継続時間	6.42 分	6.85 分	6.60 分	5.91 分
最大値	321 分	338 分	235 分	333 分
他の 3 手法の積集合	1,260	1,236	1,321	1,257
他の 3 手法の積集合と 4 つの積集合の一致率	97.3%	99.2%	92.8%	87.5%

3.6　3 σ法を用いたバーストの検出

本節では、全期間を対象として、3 σを用いたバーストの検出手法について詳述する。3 σ法では、閾値を設定、算出し、その閾値に基づきバーストの検出を行う。閾値算出は、時間 (分) ごとに行うこと、また、平日、休日で分けて算出することは先に述べた通りである。しかしこれに加えて、Twitter の利用の増加を鑑み、期間ごとにバースト閾値を変える必要がある。表 14 は、月ごとの 1 日当たりの平均ツイート数と、ユニークユーザ数を示している。表を見て分かるように、1 年間で、ツイート数及びユーザ数は増加している。2011 年末が 1,000 万ツイート前後であったのに対し、2013 年になると、1,300 万ツイートを超える。ユニークユーザ数も同様に増加している。一方、「ツイート/ユーザ」はユーザ当たりのツイート数であるが、期間中ほとんど変化していない。また、図 12、図 13 は 3 ヶ月ごとの平均ツイート数の推移を示しているが、平日、休日ともに、日が経つにつれツイート数が増えていることが分かる。特に、ツイート数の多い 19 時から 24 時は、時期による差が大きい。このことから、もし、全期間で同じ閾値を設定すると、2011 年 11 月のバーストなど、過去のバーストは検出されにくく、2013 年 1 月など新しい時期に起きたのバーストは検出されやすくなる。そこで、一定の期間を設定し、期間ごとに閾値の算出を行う。

ある期間の 1 ヶ月分の閾値を算出するためには、当該期間と、前後 1 ヶ月をデータセットに含める。表 15 に検出期間と検出に用いたデータセットの期間、及び検出されたバーストの数を纏めた。表 15 からも明らかなように、バーストの検出に用いるデータは、2011 年 11 月 16 日から 2013 年 2 月 15 日であるが、バーストの検出は、2011 年 12 月 16 日から 2013 年 1 月 15 日までの期間が対象となる。検出されたバーストは 2012 年 12 月 16 日から 2013 年 1 月 15 日までの期間が最も多く、平日で 1,053 回 (分)、休日で 1,207 回 (分) である。この期間、平日でバースト数が多いの

は、2013 年 1 月 14 日の雪によるバーストが含まれているためであると考えられる。実際、1 月 14 日は、9 時 36 分から 17 時 24 分の 469 分間バーストが継続していた。2012 年 12 月 16 日から 2013 年 1 月 15 日までの期間、休日でバースト数が多いのは、お正月が含まれているためである。実際、12 月 31 日 18 時 26 分から 1 月 1 日 6 時 56 分まで、751 分間バーストが継続していた。また、2012 年 7 月 16 日から 8 月 15 日は、平日、休日ともにバーストの検出数が多いのは、この期間、オリンピックが開催されていたためと推察される。

表 14　1 日当たりの平均ツイート数 (月ごと)

期間	ツイート数	ユニークユーザ数	ツイート数/ユーザ
2011/11/16〜12/15	10,175,501	113,154	3.0
2011/12/16〜1/15	9,959,196	111,298	2.9
2012/1/16〜2/15	10,498,452	113,942	3.0
2012/2/16〜3/12	9,953,731	127,896	2.9
2012/3/13〜4/13	10,648,164	115,669	2.9
2012/4/14〜5/14	10,265,604	120,020	2.8
2012/5/15〜6/14	11,760,679	125,677	3.0
2012/6/15〜7/14	11,900,436	129,238	3.1
2012/7/15〜8/14	12,403,368	127,130	3.1
2012/8/15〜9/14	12,194,763	130,090	3.0
2012/9/15〜10/14	12,242,603	136,425	3.0
2012/10/15〜11/16	12,210,167	133,457	2.9
2012/11/17〜12/16	12,302,317	144,700	2.8
2012/12/17〜1/16	12,921,195	145,272	2.9
2013/1/17〜2/16	13,555,443	153,905	2.9

バーストは 11,976 の時間 (分) 検出され、平日に検出されたバースト数は、6,268 回 (分) 、休日に検出されたバースト数は、4,708 回 (分) であった。また、連続したバーストを 1 回のバーストと数えた場合、バースト回数は 2,079 回であった。1 日当たりのバースト検出数が最も多いのは、2012 年 12 月 31 日で、1 日当たり、1,077 回 (分) のバーストが検出された。2 番目に多いのは、2012 年 9 月 30 日の 1 日当たり 626 回 (分) のバーストが検出された。この日は、台風が上陸した日である。3 番目に多いのは、2013 年 1 月 14 日で、541 回 (分) のバーストが検出された。この日は、東京で大雪が降った日である。

検出されたバーストについて、主なバースト要因について纏めたものが表 16 である。なお、バースト要因は、ついっぷるトレンドの HOT ワード[*13]や、バースト時にみられるツイートのテキスト

[*13] http://tr.twipple.jp/hotword/

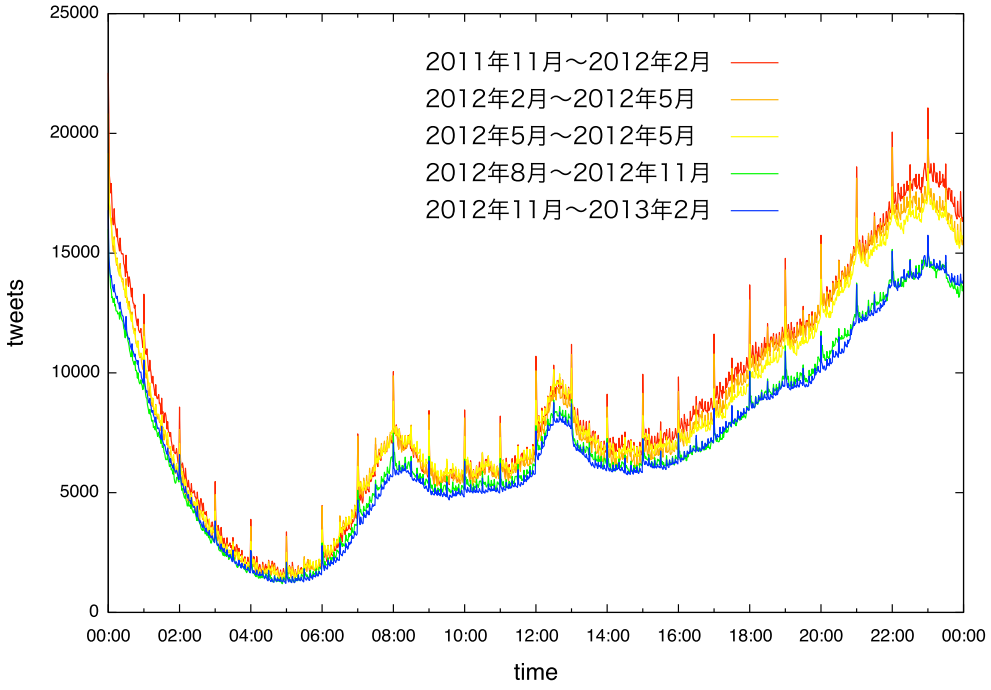

図 12　平均ツイート数の推移 (平日)

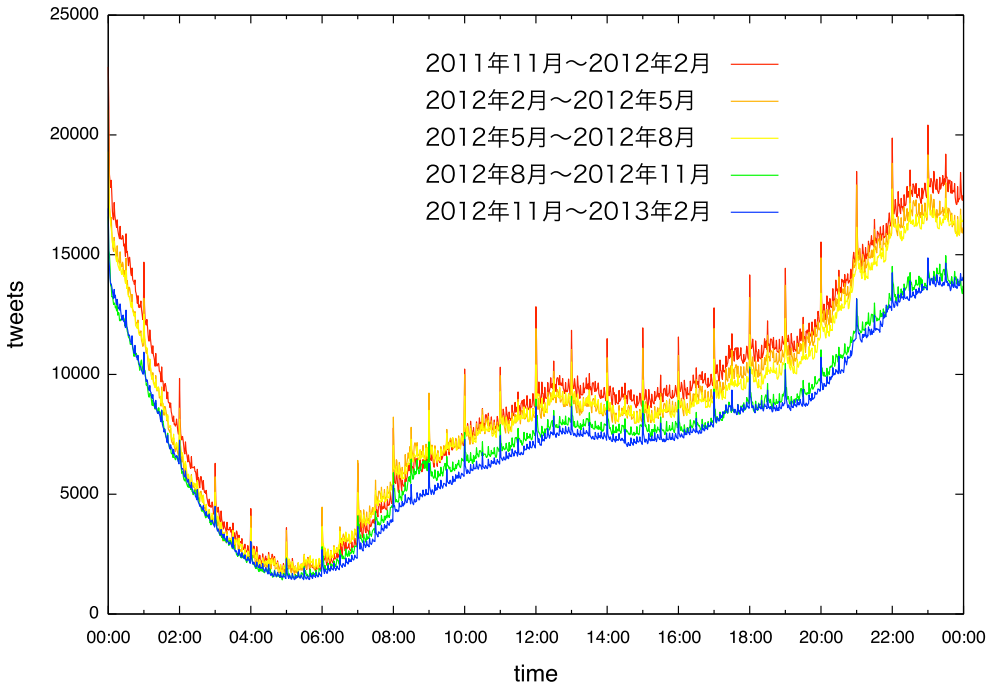

図 13　平均ツイート数の推移 (休日)

表 15　バースト検出に用いるデータセット及び検出されたバーストの数 (分)

検出期間	算出のデータセット	平日	休日
2011/12/16〜2012/1/15	2011/11/16〜2012/2/15	104	922
2012/1/16〜2012/2/15	2011/12/16〜2012/3/15	850	126
2012/2/16〜2012/3/15	2012/1/16〜2012/4/15	188	66
2012/3/16〜2012/4/15	2012/2/16〜2012/5/15	609	124
2012/4/16〜2012/5/15	2012/3/16〜2012/6/15	54	111
2012/5/16〜2012/6/15	2012/4/16〜2012/7/15	659	221
2012/6/16〜2012/7/15	2012/5/16〜2012/8/15	506	235
2012/7/16〜2012/8/15	2012/6/16〜2012/9/15	958	569
2012/8/16〜2012/9/15	2012/7/16〜2012/10/15	206	245
2012/9/16〜2012/10/15	2012/8/16〜2012/11/15	299	763
2012/10/16〜2012/11/15	2012/9/16〜2012/12/15	253	70
2012/11/16〜2012/12/15	2012/10/16〜2013/1/15	529	49
2012/12/16〜2013/1/15	2012/11/16〜2013/2/15	1,053	1,207

内容などから推定した。ついっぷるトレンドとは、ツイッター上で盛り上がっている話題／トレンドを分析して、ランキングやグラフ形式での表示するウェブサイトであり、HOT ワードは “日本国内のツイートに含まれるワード＆ハッシュタグの出現回数の急上昇度合い（＝盛り上がり度）をもとに算出”[66] されている。

‘2012 年 3 月 30 日に放映された「ルパン三世カリオストロの城」、2012 年 7 月 6 日に放映された「千と千尋の神隠し」、2012 年 12 月 5 日の「FNS 歌謡祭」などのようなテレビ番組によるバーストが複数回見られる。これに加え、大相撲 5 月場所旭天鵬優勝 (2012 年 5 月 20 日)、フェデラー優勝 (2012 年 7 月 9 日)、オリンピックの各種目 (2012 年 7 月 25 日〜8 月 12 日) などスポーツ関連のバーストが生起しやすいことが分かる。なお、これらのバーストはすべてテレビで放映されていたこと、さらに 2012 年 6 月 15 日、オウム真理教高橋容疑者逮捕によるバーストは NHK ニュース速報の 3 分後に生起していることからバースト生起へのテレビの影響が窺える。加えて、「アニソン三昧 Z (2012 年 6 月 16 日)」や「“歌う声優” 三昧 (2012 年 12 月 24 日)」はラジオで放送されていたことや、Apple 世界開発者会議 (2012 年 10 月 24 日) は公式アプリで基調講演のリアルタイム配信がなされていたことから、バースト生起と他のメディア、特に、テレビやラジオ、アプリによるリアルタイム配信など速報性の高いメディアとの関連が予想される。

一方、Twitter サーバダウンに代表されるように、他のメディアに拠らない Twitter 固有のバーストも見られる。また、災害との関連性も深く、2012 年 12 月 21 日、2012 年 4 月 29 日を始めとした地震によるバーストや、2012 年 5 月 6 日には竜巻によって、2012 年 9 月 30 日は台風によっ

てバーストが生起した。金正日死去 (2011 年 11 月 16 日)、Whitney Houston 死去 (2012 年 2 月 12 日)、森光子死去 (2012 年 11 月 14 日) のように、有名人の死去のニュースによるバーストも複数回見られた。さらに、2 月 22 日 2 時 22 分や同日 22 時 22 分、2012 年 1 月 17 日の阪神淡路大震災から 1 年や、3 月 11 日の東日本大震災から 1 年というように、バーストが生起した時間が重要性を持つことによって生起するバーストも複数回見られた。

ジャスティン・ビーバーのＴＶ出演と東京における集中豪雨によるバーストなど、バースト要因として複数の要因が絡んでいると考えられるバーストも見つかっている。

表 16　対象期間中の主なバースト要因

2011 年 12 月	2012 年 1 月	2012 年 2 月	2012 年 3 月	2012 年 4 月
[16 日] 金正日死去	[1 日] あけまして おめでとう	[12 日]Whitney Houston 死去	[11 日] 東日本大 震災から 1 年	[1 日] エイプリル フール
[18 日] クラブワールド カップ決勝	[1 日] 箱根駅伝	[22 日] 渋谷で通り魔	[30 日] プロ野球 阪神対 横浜戦	[3 日] 爆弾低気圧
[21 日] 埼玉県震度 2 の 地震	[12 日] 緊急地震速報	[22 日] 2/22 2:22 2/22 22:22	[30 日] ルパン三世カリ オストロの城	[29 日] 千葉県震度 5 弱 の地震
[25 日] 有馬記念	[17 日] 阪神大震 災から 1 年	[29 日] 大雪	[31 日] アナログ 放送終了	[29 日] 仮面ライ ダーフォーゼ
2012 年 5 月	2012 年 6 月	2012 年 7 月	2012 年 8 月	2012 年 9 月
[6 日] 竜巻	[14 日] AKB 総選挙	[3 日] 神奈川県震 度 3 の地震	[〜12 日] オリン ピック各種目	[18 日]AKB じゃ んけん大会
[20 日] 大相撲 5 月場所 旭天鵬優勝	[15 日] オウム心理教高 橋容疑者逮捕	[6 日] 千と千尋の神隠 し	[13 日] オリンピック閉 会式	[19 日] 雷雨
[21 日] 金環日食	[16 日] アニソン三昧 Z	[9 日] ウィンブル ドン Roger Federe 優勝	[23 日] 甲子園決勝選	[26 日] 自民党総裁選
[28 日] ゲリラ豪雨	[22 日]Twitter サーバダウン	[28 日] オリン ピック開会式	[30 日] 宮城県沖 震度 5 強の地震	[30 日] 台風
2012 年 10 月	2012 年 11 月	2012 年 12 月	2013 年 1 月	
[8 日] 山中伸弥 ノーベル賞受賞	[1 日] 日本シリー ズ第 6 戦	[5 日] FNS 歌謡祭	[1 日] あけまして おめでとう	
[16 日] サッカー日本代 表欧州遠征	[9 日] ヱヴァンゲリヲ ン Q	[12 日] 北朝鮮ミサイル 発射	[8 日] Nintendo Direct	
[24 日]Apple 世 界開発者会議	[14 日] 森光子死去	[16 日] 第 46 回 衆議院議員選挙	[14 日] 成人式	
[24 日] 石原都知事辞任	[28 日] ベストアーティ スト 2012	[24 日] “歌う声優” 三昧	[14 日] 大雪	

4　結果と考察

4.1　バースト時と非バースト時の比較

バースト時のユーザの投稿特性を把握するために（1）ツイートの平均文字数、(2）ツイートに占めるリツイート (RT) の比率、(3）ツイート数に占めるリプライ (@) の比率を算出し、バーストが検出されたツイートと検出されなかったツイートとを比較する。バースト時と非バースト時、全体の基本統計をそれぞれ表 17 に示す。

表 17　バースト時と非バースト時の基本統計

	全体	バースト時	非バースト時
データ数・ツイート数	5,272,178,419	135,627,887	5,136,550,532
分数 (分)	571,680	10,976 (1.92%)	560,704 (98.08%)
平均文字数（文字)	45.8	42.2	45.8
リツイート比率 (%)	8.84	9.29	8.83
リプライ (%)	39.02	33.83	39.15

バースト時は非バースト時に比べ平均文字数が少ないことが分かる。これは、バースト時は時間とツイート内容の関係性が強く、即時性が重要度を増すことによると推測される。つまり、バースト時はユーザは猶予なく情報の伝達を試み、文字数を最小限にとどめるため、本文が短いという特徴が生じる。例えば、地震時は、「ゆれた」、「jisin」、「こわい」、「キャー」といった短いツイートが随所で確認された。バースト時の文字数分布と非バースト時の文字数分布を比較したものが、図 14 である。なお、y 軸はツイート数を示しているが、分布の比較を行うため、最大値を 1、最小値を 0 とする正規化を行っている。図を見て分かるように、バースト時は、20 文字より短いツイートが非バースト時に比べ多い。特に 2-10 文字の範囲を見ると、非バースト時に比べ、バースト時のツイート数が多い。最頻値はバースト時は 20 文字、非バースト時は 22 文字であり、21 文字以上になると、20 文字以下の場合とは逆に、非バースト時の方がツイート数は多い。しかし、140 文字ツイートは、バースト時の方が多いことが分かる。

また、リプライ (@) の割合がバースト時は減少する一方で、リツイート (RT) の割合は増加する傾向にある。リツイート (RT) の比率が高いのはイベントに関する情報を不特定多数へ拡散するというユーザの意図に起因すると考えられる。例えば、リツイート (RT) の比率が高かったものとして、2012 年 3 月 14 日の地震によるバーストが挙げられる。このときの、リツイート（RT）の比率は 20.9% にのぼり、ツイート本文を確認すると “RT @zishin3255_2: ■■緊急地震速報 (第 12

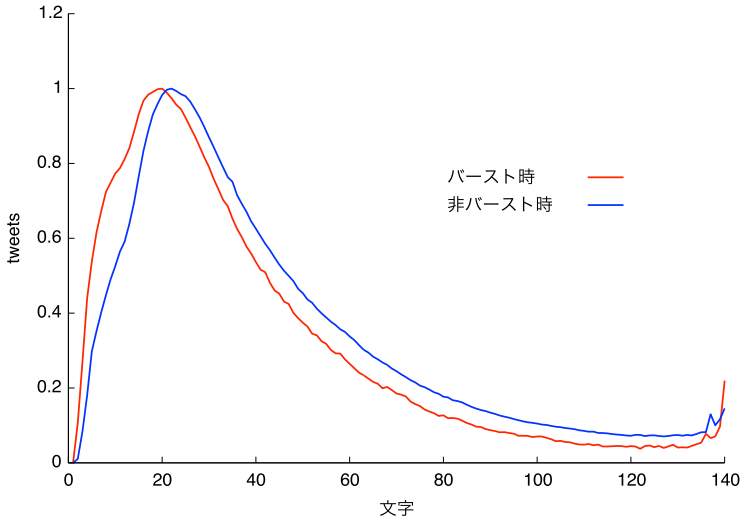

図 14　バースト時と非バースト時の文字数分布

報) ■■ 三陸沖で地震　最大震度 3 (推定) [詳細] 2012/3/14 18:08:29 発生　M7.0 深さ 10km 東京到達時刻：18:11:26 (あと約 177 秒) #緊急地震速報 ”や、“RT @NHK_PR：青森県太平洋沿岸 岩手県に津波注意報が出ています” など公共機関の防災情報のリツイート（RT）が多数確認された。このようにバースト時はユーザが情報の拡散を意図してツイートを投稿するためリツイート（RT）の比率が高く、逆に特定のユーザに向けたツイートは減少し、リプライ (@) の比率が減少すると考えられる。Twitter がコミュニケーション・ツールというよりは情報源としての性質の強いことは既に Kwak[29] により指摘されているが、バースト時は特に、情報入手、拡散ツールとしての特性が強くなることが分かる。

なお、各特徴量間の相関を表したものが表 18 である。リツイート (RT) と平均文字数の相関係数は 0.58 であり、有意水準 1% で有意である。このことから、リツイート (RT) と平均文字数には比較的高い正の相関があるといえる。図 15 は、リツイートの比率が高かった 2012 年 3 月 14 日の地震によるバースト時の文字数分布と非バースト時の文字数分布を比較したものだが、地震時は、135 文字、140 文字のツイートが非バースト時のツイート数を大幅に上回っていることが分かる。同様に、リプライ (@) と平均文字数の相関は 0.64 であり、有意水準 1% である。このことから、リプライ (@) と平均文字数にも比較的高い正の相関があると言える。

4.2　バーストの類型化

4.1 節ではバースト時は文字数が短く、リツイート（RT）比率が高く、リプライ (@) の比率が低いという特徴が示された。しかし、これらは、バースト時全体の特徴である。しかし、1 つ 1 つのバーストに着目すると、バーストを生起させたイベントによって、特徴が変わるのではないか

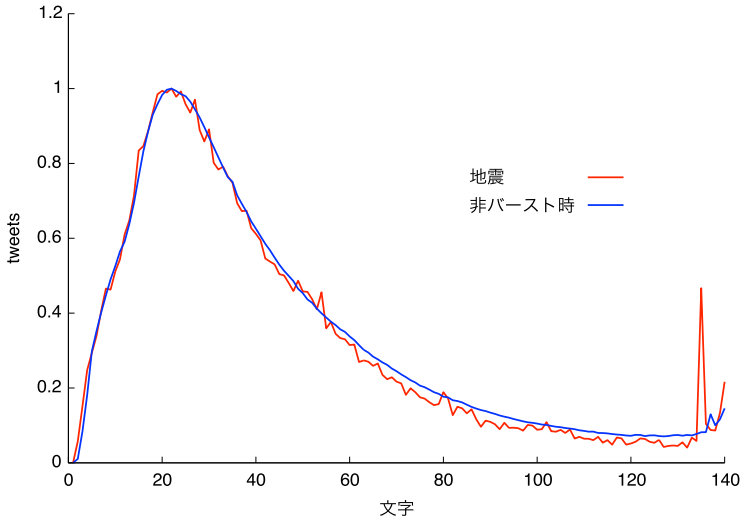

図 15　地震時の文字数分布

表 18　各特徴量間の相関

	リツイート比率	リプライ比率
平均文字数	.577**	.642**
リツイート比率		.270**

(**は有意水準 1% で有意 (両側))

という仮説が立てられる。そこで特徴量に応じて、検出されたバーストの類型化を試みる。

類型化に用いる特徴量としては、まず、4.1 節で明らかにしたバースト時の特徴量であるリプライ (@) の比率、リツイート (RT) の比率、文字数が挙げられる。これに加え、バーストの形体もイベントの性質を反映していると考えられる。例えば、図 16 を見て分かるように地震など予測不可能な現象が起こった際はイベント直後にツイート数が急激に伸び、短期間でイベント前のツイート数まで収束する。一方、金環日食のようにイベント中に最も盛り上がる時間が分かっている場合はその頂点に向かってツイート数が徐々に増加し、頂点を過ぎると緩やかに減少していく（図 17)。図 18 はサッカー日本対豪州戦が開催された 6 月 12 日のツイート数の推移を示したものである。試合開始時間である 19 時より前から通常時に比べるとツイート数が多く、試合終了時間の 20 時 50 分以降もしばらくは通常と比較してツイート数が多い。さらにその中でもゴール時や試合終了時はツイート数が急激に増加していることが分かる。特に、日本対オーストラリア戦において本田選手がフリーキックを蹴る前にホイッスルが鳴らされたことによる不満が、試合終了時のツイート数の増加に拍車をかけたことが予測される。最後に、図 19 に示した爆弾低気圧によるバーストは平均との差はそれほど大きくないものの、長時間バーストが続いている。このように文字数、リツ

イート (RT) 比率、リプライ (@) 比率に加え、ツイート数の増減傾向、バーストの高さや長さがバーストを生起させたイベントに対する Twitter ユーザの認識を表現していると考えられる。

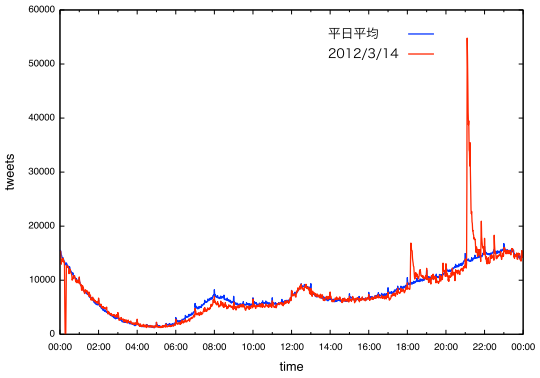

図 16　地震によるバースト

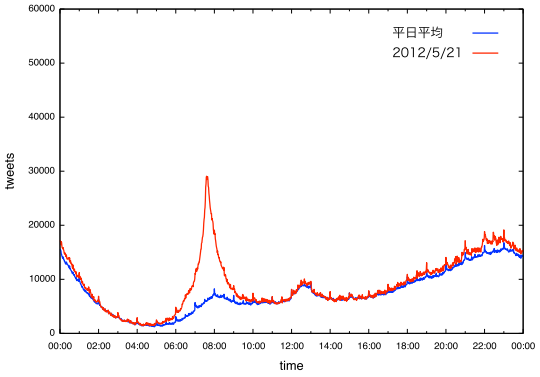

図 17　金環日食によるバースト

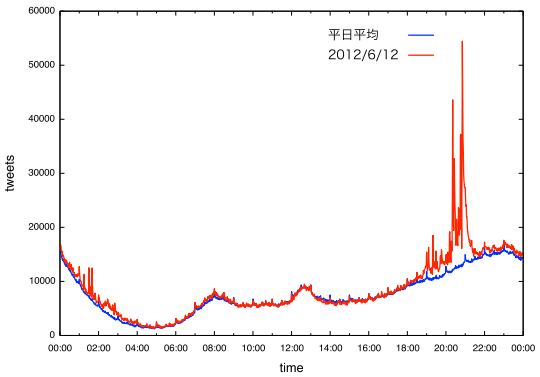

図 18　サッカー日本-豪州戦によるバースト

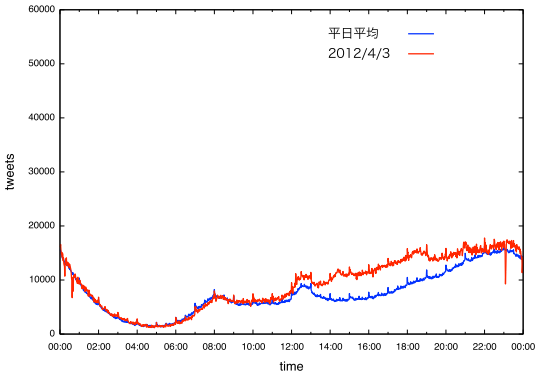

図 19　爆弾低気圧によるバースト

以上のことからバーストの形体、及び、投稿特徴を含めた形質がイベントの性質を表しているという仮説が導かれる。そこでバーストを特徴量により各バーストを類型化し、各々の類型についてイベント性質を把握することを試みる。特徴量はバーストの形体を表す (1) バーストの継続時間、(2) 閾値との差、さらに、バースト時の特徴的な傾向である (5) ツイート本文の平均文字数、 (3) リプライ（@）の比率、(4) リツイート（RT）の比率とする。バーストの継続時間とはツイートがバースト閾値を連続して越えた場合の最初に越えた時刻からバースト閾値を下回るまでの合計時間を指す。連続していない場合は、継続期間は 1 分間である。バーストが継続している場合、その継続期間全体を一つのバースト現象として算出すると、調査期間中 2,079 回のバースト現象が観測された。この 2,079 回のバーストについて基本統計をまとめたものが表 19 である。なお、ツイート数は 1 分あたりの平均ツイート数である。

表 19　バーストの基本統計（継続するバーストを一つのバーストとした場合）

	ツイート数	継続分数	閾値との差	平均文字数	@比率	RT 比率
平均	10823.75	5.33	1336.47	42.50	0.34	0.08
標準偏差	5729.91	26.41	4028.07	4.36	0.06	0.03
最大値	29978.29	751	97887.95	54.65	0.60	0.40
中央値	9664	2	261.28	43.05	0.35	0.08
最小値	1256	1	0.00	21.54	0.12	0.00

次に、特徴量に応じてバーストのクラスタリングを試みた。ここで、継続しているバーストを一つのバーストとすると、同じ要因によるバーストであっても 1 度閾値を下回った場合は別のバーストとして扱われることが懸念される。また、同じ要因によるバーストでも時間によって特徴量が変わっていくことがある。例えば、災害バーストにおいて、災害直後は被災者が現状をツイートすることによってバーストが生じるが、時間が経過するに従って、災害情報や他のユーザのツイート

のリツイート (RT) によってバーストが起こる、ということが考えられる。したがって特徴量によって分類を試みる場合、継続期間をまとめて分析するのは不適当だと考えられる。そこで、継続時間については特徴量の一つとして分析するにとどめ、クラスタに分類する際は 1 分ごとのバースト（10,976 回 (分)）の特徴量から類型化する。類似度にはユークリッド距離を、結合アルゴリズムは Ward 法を用いる。ここで、ユークリッド距離では、値の大きさがそのまま距離に反映されてしまう。このため、比率であり、最大値が 1 を超えることがないリツイート (RT) 比率やリプライ (@) 比率に比べ、継続分数や閾値との差、平均文字数の影響が強く出すぎることが懸念される。そこで、継続分数、閾値との差、平均文字数については、平均が 0、分散が 1 になるように正規化したデータを投入した。10,976 回 (分) のバーストについて、基本統計を纏めたものが表 20 であり、クラスタ分析の結果を表 21 に示す。なお、個数は、そのクラスタに含まれるバーストの個数を示している。

表 20　バーストの基本統計

	ツイート数	継続分数	閾値との差	平均文字数	@比率	RT 比率
平均	12356.77	137.46	1908.08	42.42	0.34	0.09
標準偏差	7014.44	221.52	3537.09	5.02	0.07	0.05
最大値	126406	751	97887.95	58.67	0.62	0.47
中央値	11198	21	734.54	43.52	0.35	0.09
最小値	1237	1	0.00	13.75	0.03	0.00

表 21　クラスタ分析の結果

	継続時間	閾値との差	平均文字数	@比率	RT 比率	個数
1. 小さなイベント型	21.32	567.44	44.87	38.28	9.42	4,356
2. 既知イベントピーク型	549.91	4110.08	43.41	36.10	10.95	2,125
3. 既知イベント準備期間型	54.76	1245.37	40.35	30.86	6.99	3,116
4. 突発的イベント型	51.57	5366.71	31.36	20.07	5.15	870
5. 情報拡散型	62.40	2333.45	48.86	27.94	22.21	509

クラスタごとに特徴量を見ると、第 1 クラスタは、すべてのクラスタの中で最も継続時間が短く、閾値との差も少ない。このことから、第 1 クラスタは、「小さなイベント型」と言える。ここには、震度 1 の地震や、ゲリラ豪雨などのように、限られた地域の人が影響を受けたイベントによるバーストが含まれていた。なお、含まれるバーストの個数は 4,356 件と、第 1 クラスタに分類されるバーストが最も多かった。第 2 クラスタは、閾値との差が大きく、継続時間も長いことから、「既知イベントピーク型」といえる。ここには、多くの人々が準備し、ツイートを行うイベントで

ある「あけましておめでとう」バーストや金環日食の、「金環」が起きている間のバーストなどが含まれていた。第 3 クラスタは、第 1 クラスタに比べ継続時間が長く、第 2 クラスタに比べ継続時間が短い。閾値との差も同様に、第 1 クラスタに比べ長く、第 2 クラスタに比べ短い。実際に、ここに含まれるイベントを見てみると、金環日食が始まる前のバーストなどがここに含まれていた。以上のことから第 3 クラスタは、「既知イベント準備期間型」と言える。第 4 クラスタは、閾値との差が最も大きい。しかしながら、継続時間は比較的短い。それゆえ、第 4 クラスタは、「突発的イベント型」と言える。ロンドンオリンピックで勝敗が決定した瞬間や、サッカーのゴール時に起きたバーストがここに含まれていた。第 5 クラスタは、リツイート (RT) の比率が最も高いことから、「情報拡散型」と言える。リツイート (RT) 比率が高いことから、平均文字数も長いという特徴も見られる。渋谷で無差別殺人者が現れたことを知らせるバーストや、高橋容疑者逮捕によるバーストがここに含まれていた。

以上のことから、各バーストは、「小さなイベント型」、「既知イベントピーク型」、「既知イベント準備期間型」、「突発的イベント型」、「情報拡散型」の 5 つに分類できることが分かった。

4.3 地震バーストに影響を与える要因

Twitter は即時性といった特性をもち、140 文字という字数制限もあいまってユーザがツイートした直後に他のユーザのタイムラインにツイートを表示することができる。この特性に起因するリアルタイム性が、速報性の高い情報の提供を可能にしている。榊ら [61] は、ソーシャルメディアが現れる以前は、特にリアルタイム性の高い情報についてはテレビやラジオが大きな役割を担っていたが、最近では既存のメディアに加えて Twitter を利用して情報収集する人が増えていること指摘している。このようなリアルタイムな情報を提供するという特徴から、Twitter から実社会の状態を観測する研究は、少なからず行われている。例えば、ツイートデータから実社会を観測し、そこから未来の予測を行う研究として、Bollen[62] は株価の予測を、Asur[63] は映画の興行収入を、Tumasjan[64] は選挙結果をツイートデータから予測している。

実社会を観測するという面で、特に注目されているのが、災害情報の提供である。これまで、マスメディアによる報道は、報道関係者が情報を収集し、提供するという構造であった。しかし、ソーシャルメディアでは、災害を経験したユーザ自身が情報を提供することから、即座に情報が提供される。Twitter と災害との関連性を明らかにすることは、災害情報の迅速な提供に資すると考えられる。Sakaki ら [65] は位置情報とツイート内容から地震情報の推定、提供を試みた。Sakaki の手法では、震度 3 以上の地震は 96%、震度 4 以上では 100% 地震の発生を観測したことが報告されている。しかし、バーストと災害の関係と言った観点からの分析はこれまでほとんど行われていない。災害時はバーストが起こりやすく、調査期間中も台風、豪雨、地震等災害に起因するバーストは複数回観測された。なかでも地震が原因と考えられるバーストは 100 回を超える。そこで、地震によるバーストとそれに影響を与える災害の要素を検討する。

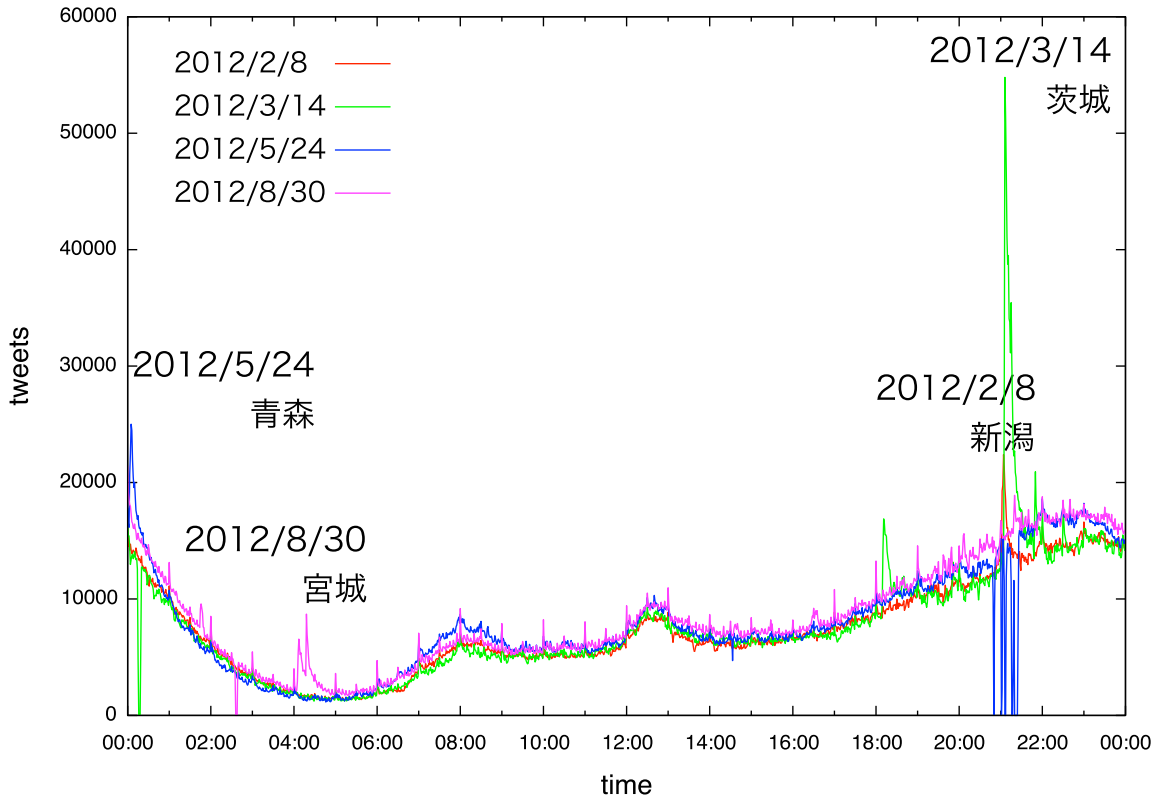

図 20　震度 5 以上の地震時のツイート分布

地震直後のバーストは揺れに反応したユーザがツイートすることによって発生するため、バーストの有無に影響を与えられる要因としてまず震度が挙げられる。これに加え、都心からの距離もバースト発生の有無に関わる要因の一つと考えられる。これは、都心に近いほど Twitter のユーザ数が多いこと、そして、近い場所の出来事に人々は敏感に反応することを考慮したためである。人々の知覚原理として、自分たちから近い場所の出来事については敏感に反応し、遠いところではそれほど敏感に反応されないという傾向がある。例えば、日本国外の事故・事件では日本人が含まれていたかどうかが必ず報道されるように、自分たちに近い場所で起こっているか否か、及び、自分たちの同胞であるか否かによって、人々の関心は異なる。このことから、都心で起きた事件には、多くの人々が関心を持つと考えられ、多くの情報発信そして情報の拡散が行われる。しかしながら、一方で、この傾向により、過疎地や遠隔地の悲劇を過小評価する危険性も孕む。

図 20 は震度 5 強の地震が起きたツイート分布に最大震度を記録した都道府県名および東京からの距離を記載したものだが、同じ震度でもツイート数に差があることが分かる。図 20 では 2012 年 3 月 4 日に茨城で震度 5 強の地震が起きた場合が、最もツイート数が大きい。また、図 21 は 2012 年 5 月 24 日の青森県東北町で最大震度 5 強を記録した地震が生起した日のツイート数と、2012 年 5 月 29 日に東京都渋谷区で最大震度 4 を記録した日のツイート数を比較したものだが、青森県で

震度 5 強を記録した地震より、東京都で震度 4 の地震を記録した場合の方がツイート数が多く、これらは都心からの距離によると推測される。また、後述するように、本研究では都心の代表地点の一つを東京都庁とするが、東京都庁から最も近い地震観測点である東京都新宿区歌舞伎町観測点では、調査期間中、震度 1 の地震が 36 回、震度 2 の地震が 11 回観測された。震度 3 の地震が 11 回、合計 50 の地震が観測された。このうちバーストが検出されたものは 46 回にのぼり、都心が揺れている場合は高い確率でバーストが検出されることが分かる。

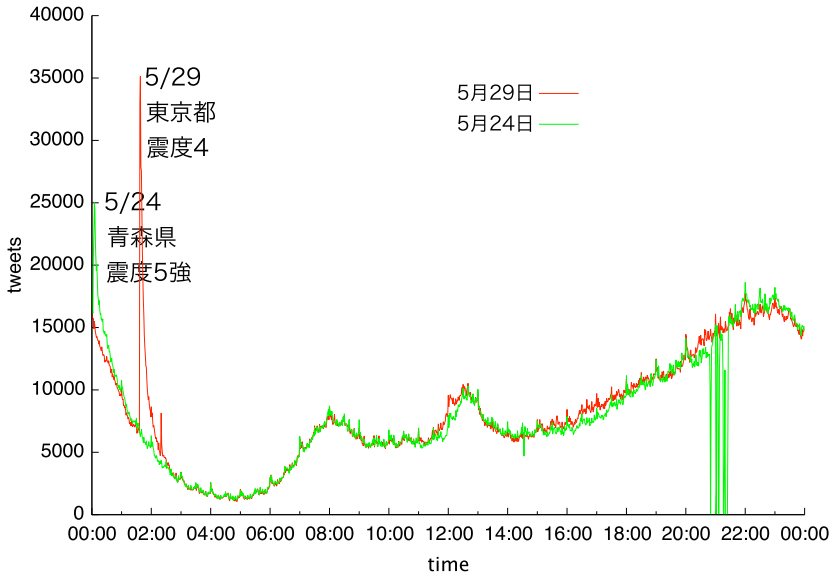

図 21　青森県と東京都の地震の比較

そこで、震度、都心からの距離がバーストの生起に影響を与えるという仮説を検証する。調査対象期間中に日本国内で最大震度 3 以上を記録した 341 回の地震について、(1) 発生時刻、(2) 震源地、(3) マグニチュード、(4) 最大震度、(5) 最大震度を記録した市町村、(6) 最大震度を記録した市町村から都心までの距離についてそれぞれデータを収集した。調査対象期間中に発生した地震とその震度、最大震度を記録した市町村は日本気象協会の提供する地震情報[*14]を参照する。本研究では、都心の代表地点を東京とした場合と、都心の代表地点を三大都市である東京、大阪、名古屋とした場合での結果の比較も行う。都心の代表地点を東京とした場合、最大震度と最大震度を記録した市町村から東京都庁までの距離を算出する。最大震度を記録した場所が複数ある場合は、東京都庁から最も近い市町村との距離を算出する。

都心の代表地点を東京、大阪、名古屋とした場合、東京都庁、大阪市役所、名古屋市役所のうち、最大震度を記録した市町村の役場からもっとも近い場所を選択し、2 点間の距離を算出する。最大震度を記録した場所が複数ある場合は、最大震度と都心からの距離が最も短い都心と最大震度を記

[*14] http://bousai.tenki.jp/bousai/earthquake/

録した市町村の組み合わせを選択する。距離の算出には Google Maps API[*15]を使用した。

地震後にバーストが生起したかどうかについては、各地震一つ一つに、○ (バースト生起)、△ (判断不可能)、× (バースト不生起) のいずれかのラベルをつける。なお、○、△、×の条件を表 22 に纏めた。地震生起後 3 分以内にバーストが生起していない場合は、バースト不生起とする。これは、P 波と S 波では 1 分ほどのタイムラグが生じることと地震が生起してから伝わるまでの猶予時間が最大 90 秒であることを考慮したためである。なお、調査期間中、地震生起後 4 分から 10 分の間にバーストが生じることは無かった。地震前 5 分以内に違うバーストが生起していた場合は、地震後にバーストが起きていてもそれが地震によるものか否かの判断がつかないため「△」のラベルを付与し、分析のためのデータセットからは除外する。また、地震生起の前 1 分以内にバーストが生起している場合は、中身を確認し、「地震」や「ゆれ」などの記述が見られれば「○」とする。例えば、2012 年 7 月 13 日 11:32 に神奈川県で最大震度 4 を記録した地震は、11:31 からバーストが生起しており、ツイート本文を見ると「地震」、「ゆらゆら」などの記述が見られたため、「○」とした。震度と都心からの距離あたりのバースト検知率を示したものが、表 23、24 である。まず、都心を東京とした表 23 から見ると、震度 3 の場合は都心からの距離が近いほどバースト検知率が高く、都心からの距離とバースト生起の有無との関連が窺える。他方、震度 5 を越えると場所に拘らず高い確率でバーストが検出されていることが分かる。 これは、震度とバースト生起の関連性を示唆する。また、都心の代表地点を東京とした表 23 と、代表地点を東京、大阪、名古屋とした 24 を比較すると、大阪と名古屋を含めた場合は 100km 以内の検知率が著しく低下している。特に、震度 4 の場合は、検知率が半分まで低下している。このとから、東京では揺れを感じるとバーストが起きやすいのに対し、大阪や名古屋の付近で生起した場合はバーストが生起しにくいと言える。

これは、東京の人口は大阪、名古屋と比較すると極めて多いことが原因に挙げられる。また、東日本大震災を経験した人々は、その体験から地震に対して敏感に反応を示し、これがツイートにも反映されたことが原因であると考えられる。

表 22 各ラベルの条件

ラベル	条件	検出数
×	地震生起後の 3 分以内にバーストが検出されていない	224
△	地震生起後 3 分以内にバーストが検出されているが、 地震前の 5 分以内に違うバーストが生起していた場合	11
○	地震生起後 3 分以内にバーストが検出されており、 地震前の 5 分以内に違うバーストが生起していない場合	106

次に、「最大震度」と「最大震度を記録した市町村から都心までの距離の逆数」を独立変数とし、

[*15] https://developers.google.com/maps/

表 23 震度、都心からの距離によるバースト検知率 (都心を東京からの距離とした場合)

都心からの距離	震度 3	震度 4	震度 5 以上
100km 以内	25/38(65.8%)	16/16(100%)	2/2(100%)
100-200km	14/86(16.3%)	12/21(57.1%)	4/6(66.7%)
200-300km	4/27(14.8%)	3/3(100%)	4/4(100%)
300km 以上	10/98(10.2%)	9/26(34.6%)	7/7(100%)

表 24 震度、都心からの距離によるバースト検知率 (都心を三大都市とした場合)

都心からの距離	震度 3	震度 4	震度 5 以上
100km 以内	25/42(59.5%)	16/32(50%)	2/2(100%)
100-200km	15/87(17.2%)	12/21(57.1%)	3/5(66.7%)
200-300km	5/29(17.2%)	3/3(100%)	3/3(100%)
300km 以上	8/91(8.79%)	9/26(34.6%)	5/5(100%)

バースト生起の有無を目的変数としたロジスティック回帰分析を行った。分析には R2.15.1 の glm 関数を用いた。

ロジスティック回帰推定の結果を表 25、表 26、表 27 に示した。表 25 はモデルの適合度を表す。都心を東京とした場合、McFadden の ρ は 0.26[*16]、Cox-Snell の R^2 は 0.366、Nagelkerke の R^2 は 0.488[*17]といずれも高い値を示している。また、推定された回帰式に基づく判別率は 80.5% に上った。一方、都心を三大都市とした場合は、McFadden の ρ は 0.036、R^2 は 0.159、Nagelkerke の R^2 は 0.212 といずれも東京を都心とした場合を下回る。また、判別率も 68.9% と、10% 以上も低下している。これは、上述したように、名古屋や大阪からの距離が近いにも関わらず、バーストが生起しなかったことによると考えられる。

都心を東京とした場合、各々の独立変数の偏回帰係数はいずれも有意水準 1% で有意であり、最大震度と都心からの距離の短さがバースト生起に影響を与えていることが分かった。特に、Wald 統計量の値などから最大震度よりも都心から被災地までの距離の短さがバースト生起により強い影響を及ぼすことが明らかになった。なお、独立変数間の相関係数は 0.1 未満であり、各々の独立変数の分散拡大要因 (VIF) はいずれも 2 を下回っていることから，多重共線性の存在も認められなかった。しかしながらこのことは人口の少ない地域における災害をメディアでは矮小化してしまうという危険性を示唆しているとも言える。

[*16] McFadden の ρ と最小二乗法の R^2 の間ではおおよそ ρ[0.1、0.2、0.3、0.4、0.5] = R^2[0.3、0.5、0.6、0.8、0.9] という対立関係が成り立つ

[*17] NagelkerkeR^2 は Cox-SnellR^2 を 0〜1 の値をとるように正規化したもの

表 25　モデルの適合度

当てはまりの尺度	都心を東京とした場合	都心を三大都市とした場合
-2 対数尤度	305.365	398.015
McFadden's ρ	0.260	0.036
Cox-Snell's R^2	0.366	0.159
Nagelkerke's R^2	0.488	0.212
判別率	80.5%	68.9%

表 26　ロジスティック回帰の結果 (都心を東京とした場合)

	偏回帰係数	標準誤差	Wald	有意確率	オッズ比
最大震度	5.228	0.703	56.657	5.19E-14**	197.98
都心からの距離の逆数	1.354	0.163	68.872	2E-16**	3.87

表 27　ロジスティック回帰の結果 (都心を三大都市とした場合)

	偏回帰係数	標準誤差	Wald	有意確率	オッズ比
最大震度	1.794	0.474	13.707	2.14E-4**	5.7772
都心からの距離の逆数	0.528	0.163	68.872	1.24E-6**	1.6958

4.4　バースト時に見られる感情

前述の通り、調査期間中、地震によるバーストは複数回見られた。また、サッカーによるバーストやオリンピックによるバーストなど、スポーツイベントによるバーストも少なくない。これらのバーストは感情と強く結びついていると考えられる。例えば、地震によるバーストは、ユーザが「怖い」という感情を持つことによりツイートされると考えられる。また、スポーツによるバーストは、応援しているチームが試合に勝って「嬉しい」という感情や、負けて「哀しい」という感情により引き起こされていると考えられる。そこで、それぞれのバーストについて、どのような感情によって生起しているのかを調査し、どのような感情をユーザが感じるとバーストが起きやすいのかを明らかにする。

Twitter における感情の分析としては、ツイートの中にどのような感情が含まれているかを明らかにした Bollen[62] の研究がある。Bollen は、2008 年 8 月 1 日から 12 月 20 日までのパブリッ

クタイムラインのデータ 9,664,952 ツイートに対し、Profile of Mode States (POMS) を用いて感情の分析を行った。POMS は、単語ベースの感情評価手法であり、「緊張: tension」、「憂鬱: depression」、「怒り: anger」、「活気: vigor」、「疲労: fatigue」、「困惑: confusion」の 6 種類のラベル付けを行うことができる。Bollen は POMS を用いて、アメリカ大統領選や感謝祭 (Thanksgiving Day) の前後で感情がどのように変化するかについて分析を行うとともに、DJIA (ダウ平均, Dow Jones Industrial Average) や WTI (West Texas Intermediate) の推移と感情の分布を比較し、これらが長期的な感情の変化に影響を及ぼすことを明らかにした。Poblete[31] は、英語とスペイン語を対象に、「幸せ: happiness」の感情の分析を行い、国ごとの比較を行っている。分析には、特定の単語に対し、「幸せ: happy」から「不幸せ: unhappy」の範囲で幸せのレベルを測定する Dodds ら [67] の手法を用いている。分析の結果、英語、スペイン語ともにオーストラリアの幸せレベルがもっとも高かったことなどを示した。

また、日本語のテキスト文を対象とした研究としては、三浦 [68] が、東日本大震災発生 30 分後から 40 時間分のツイート、469,504 件に対し、表出する感情の分析を行っている。データセットを 2 時間ごと、20 区間に分割し分析を行ったところ、地震直後の感情反応として最も多く表出されたのは「不安」の感情であるが、これは時間の経過に従い減退することなどを明らかにした。

ツイートに見られる感情の推定には、中村 [69] の『感情表現辞典』を参考にする。この辞典には、「喜、怒、哀、怖、恥、好、厭、昂、安、驚」の 10 種類の感情とその感情を表す言葉が収録されている。表 28 に各感情とそれを表す言葉として主なものを示した。この辞典は、日本語テキストの中から感情の推定を行う際に、一般的に用いられる辞典であり、テキストマイニングやシステムの開発から言語学、日本語教育や障害者教育に至るまで様々な分野での研究で用いられている [70][71][74][73][75]。ウェブ領域を対象にしたものについても、オンライン書評を対象とした原田ら [76] の研究、ブログを対象とした青木ら [77] の研究などで『感情表現辞典』が用いられている。Twitter を対象に『感情表現辞典』を用いた研究としては橋本ら [78] の研究がある。橋本らは、Twitter のテキストデータを用い、『感情表現辞典』をもとに抽出した感情などから、特定の商品やサービスに対する評判を明らかにするシステムの開発を行っている。また、先に述べた東日本大震災直後の Twitter 上の感情を分析した三浦 [68] も、この『感情表現辞典』を参考に感情の抽出を行っている。

『感情表現辞典』に含まれる感情語は、2 つ以上の感情に含まれる場合もある。例えば、「屈辱」は「厭」と「恥」の両方に含まれる。また、「泣く」は「喜」、「哀」、「昂」の 3 つの感情に含まれている。このような単語は、どちらの感情であるか判断ができないため、ノイズとして感情語のデータセットからは除外する。また、「昂」の項目には「ゆらゆら」、「揺れ」などの感情語が含まれている。しかし、先述したように、バーストは地震によって生起する場合も少なくなく、地震によるバーストが「昂」と推定されてしまうことが危惧される。そこで、地震と関連すると考えられる「ゆらゆら」、「揺れ」、「ゆらり」、「ぐらぐら」の 4 つの言葉を除去した「昂」の感情語データセット

も作成し、結果の比較を行う。なお、地震に関連すると考えられる 4 つの単語を地震語とし、以下の表では地震語を除去した昂のデータセットを「昂 (地震語無し)」と示す。

表 28　感情語の例

感情	感情語の例
喜	喜び、喜ぶ、嬉しい、楽しい、面白い、わくわく、満足、幸せ
怒	怒り、腹立つ、激怒、ぷんぷん、むかつく、不愉快、不満
哀	哀しさ、悲しさ、悲しい、泣ける、淋しい、虚しい、憂い
怖	不気味、怖い、恐ろしい、ぞっと、がたがた、ぶるぶる、不安、殺気
恥	恥、恥ずかしい、恥ずかしさ、照れる、赤らむ、屈辱、 赤面
好	友情、愛する、恋しい、愛しい、惚れ惚れ、憧れ、好き、愛着
厭	不快、嫌気、厭う、呪う、大嫌い、憎たらしい、妬み、鬱陶しい、辛い
昂	昂り、焦る、苛立つ、激情、ときめき、感極まる、感動、感心
昂 (地震語無し)	「昂」から、「ゆらゆら、揺れ、ゆらり、ぐらぐら」を除外したもの
安	ほっと、安らか、安らぐ、和やか、冷静、落ち着く、安楽
驚	驚く、驚き、ぎょっとする、ショック、愕然、驚愕、おろおろ、意外

バースト時と非バースト時を比較し、ツイート本文に含まれる各感情の割合が非バースト時より高い場合、その感情とバーストが強い関係を持つと見なし、バーストに当該感情のラベル付けを行う。このため、まず非バースト時の各感情の比率を求める必要がある。非バースト時の感情比率は、検出されたバーストの数と同じ分数 (10,976 分) をランダムに抽出し、そこに含まれる各感情の比率の平均を求める。ランダムな時間の抽出には、エクセルのランダム関数を用いた。このようにして求めた非バースト時の感情比率を表 29 に示す。なお、ツイート数とは、その感情が含まれているツイートの数であり、これを 10,976 分に見られた全ツイート数 570,161 ツイートで除した値が比率 (%) である。

これを見て分かるように、非バースト時のツイートには、「好」が 6.28%、「喜」が 5.31%、「安」が 3.06% というように、ポジティブな感情語が多く含まれていると言える。また、地震語を除外していない「昂」が 0.72% (622,664 ツイート)、地震語を除外した「昂」が 0.66% (570,161 ツイート) であることにより、他の「昂」に関わる感情語を含まず、地震語のみが含まれているツイートが 0.6% (52,503 ツイート) 含まれていたことが分かる。

先述したように、バーストに含まれる各感情比率が表 29 に示す非バースト時を上回った場合、その感情と強い関係があるとみなし、各バーストに感情のラベル付けを行った。例えば、表 30 には非バースト時の感情比率、新潟県佐渡市で最大震度 5 強の地震が起きた 2012 年 2 月 8 日 21 時 10 分の感情比率、オリンピックの体操競技で内村選手の金メダルが確定した 2012 年 8 月 2 日 2 時 52 分の感情比率を示している。なお、ここでの「昂」は、スペースの関係上、地震語無しのものの

表 29　非バースト時の各感情比率

感情	比率 (%)	ツイート数
喜	5.31	4,605,722
怒	0.42	361,820
哀	1.76	1,522,240
怖	0.60	212,327
恥	0.24	517,989
好	6.28	5,440,142
嫌	2.73	2,367,279
昂	0.72	622,664
昂 (地震語無し)	0.66	570,161
安	3.06	2,652,041
驚	0.54	470,032

みを記載している。表 30 を見ると、2 月 8 日の地震時のバーストでは、「怖」と「驚」が非バースト時の感情比率を上回っているため、「怖」と「驚」の感情をラベル付けする。また、8 月 2 日の金メダル確定時のバーストでは、「喜」と「昂」が非バースト時の感情比率を上回っているため、「喜」と「昂」の感情をラベル付けした。

以下に結果を示していくが、まず、各感情の共起率を示す。「共起」とは、同じバースト中で 2 つの異なる感情が生起していることを指す。表 30 では、地震のバーストでは「怖」と「驚」が共起しており、金メダル確定のバーストでは「喜」と「昂」が共起していると言える。

表 30　各感情比率の例

	喜	怒	哀	怖	恥	好	厭	昂	安	驚
非バースト	5.31%	0.42%	1.76%	0.60%	0.24%	6.28%	2.73%	0.66%	3.06%	0.54%
地震	3.97%	0.26%	1.35%	**1.02%**	0.15%	5.64%	2.67%	0.53%	2.54%	**0.56%**
金メダル	**5.59%**	0.07%	1.64%	0.18%	0.07%	1.72%	1.80%	**2.33%**	0.89%	0.10%

「共起率」とは、その 2 つの感情語を含んだすべてのバーストの中で、2 つの感情が共起しているものの割合であり、次式で表せる。P_{AB} は A と B の一致率であり、A_{em} は A の感情を持つバーストの集合、B_{em} は B の感情を持つバーストの集合である。また、$A_{em} \cap B_{em}$ は A_{em} と B_{em} の積集合を $A_{em} \cup B_{em}$ は A_{em} と B_{em} の和集合を指す。

$$P_{AB} = \frac{A_{em} \cap B_{em}}{A_{em} \cup B_{em}} \tag{18}$$

まず、地震語を除外していない「昂」の他の感情との共起率と、地震語を除外した「昂」と他の感情との共起率を表 31 に示した。地震語有りの場合は、「怖」との共起率が 31.36% と最も高く、これは、「揺れ」、「ぐらぐら」など地震時にみられる単語が「昂」の感情と誤って推定されたためと考えられる。実際、地震語無しの場合は「怖」との共起率は 24.13% と、地震語を除外していない場合の 31.36% と比べると大幅に減少している。地震語無しの「昂」と共起率が最も高いのは「喜」の 28.92% であり、地震語を除外した場合の方が、「昂」の感情が正しく推定されていることが示唆される。そこで、以下の分析では、地震語を除外した「昂」を用いる。

表 31　昂と各感情の共起率

共起感情	昂 (地震語有り) との共起率 (%)	昂 (地震語無し) との共起率 (%)
喜	21.91	28.92
怒	24.78	16.99
哀	27.07	26.12
怖	31.36	24.13
恥	19.59	20.08
好	22.83	24.98
厭	20.87	23.66
安	15.70	20.07
驚	20.09	20.51

表 28 に各感情間の共起率を示した。もっとも共起率が高いのは、「喜」と「安」の 50.07% である。さらに、共起率が 25% を超えるのは、「喜」と「安」の 26.13%、「哀」と「怖」の 26.16%、「哀」と「厭」の 27.40%、「昂」と「喜」の 28.92%、「昂」と「哀」の 26.12% である。「昂」は感動する、ときめく、などのポジティブな言葉と、焦る、苛立つなどのネガティブな言葉を包含する。このため「昂」に関しては考察し難いが、「喜」と「安」などポジティブワード同士、そして「哀」と「怖」、「哀」と「厭」のネガティブワード同士が共起していることから、感情推定がある程度の精度を持つと言える。

次に、すべてのバースト 10,976 件にラベルづけられた各感情の個数を多い順に表 33 に示す。最も多いのが「昂」であり、バーストの 40.6% が「昂」の感情を含むことが分かる。次に多いのが「怖」の 3,766 件であり、恐怖を感じたことによるバーストが多いことが分かる。これは、地震によるバーストが多いという結果と整合している。3 番目に多いのが「哀」であり、ユーザが哀しさ

表 32　感情の共起率 (%)

	怒	哀	怖	恥	好	厭	昂	安	驚
喜	6.63	19.25	9.10	13.93	18.50	14.76	21.91	53.07	15.11
怒		19.76	26.13	17.39	15.43	18.25	24.78	6.41	18.76
哀			26.16	22.63	17.67	27.40	27.07	13.83	22.06
怖				21.40	20.38	21.05	31.36	9.75	24.42
恥					18.21	24.46	19.59	11.97	19.93
好						21.57	22.83	15.64	20.57
厭							20.87	10.20	22.75
昂								15.70	20.09
安									13.91

ことによるバーストが多いことが確認された。

表 33　各感情のバースト数

順位	感情	バースト数
1	昂	4,460
2	怖	3,766
3	哀	3,381
4	喜	3,092
5	驚	2,721
6	怒	2,679
7	恥	2,526
8	好	2,519
9	厭	2,491
10	安	2,414

次に、平均より感情語の比率が高かった場合の閾値との差の平均を表したものが表 34 である。表 34 は値の大きいものから順に示しており、最も多いのが、「好」の 1978.58、次に多いのが「昂」の 1836.23、3 番目に多いのが「喜」の 1535.46 である。閾値との差の平均が高いものはポジティブな感情であることが分かる。以上のことから、バーストの多くがネガティブな感情によるものだが、バーストした際に、ツイート数が多くなるのは、ポジティブな感情により生起したバーストで

表 34 閾値との差の平均

順位	感情	閾値との差の平均
1	好	1978.58
2	昂	1836.23
3	喜	1535.46
4	哀	1479.09
5	怒	1400.16
6	怖	1338.95
7	恥	1290.21
8	厭	1193.83
9	驚	1104.15
10	安	981.69

あると言える。

5　結論

本研究では、Twitter 上でのユーザが種々の社会的事象をどのように捉え、伝達しているのかというように、ユーザの情報行動を明らかにすること、さらにこれにより、Twitter のメディア特性を描写することを目的として、Twitter におけるバーストの分析を行った。具体的には、どのようにバーストの検出を行うのが妥当であるか、どのようなイベントによってバーストが生起するのか、バースト時はどのような投稿特徴があるのかについて分析と考察を行った。

バーストの検出手法については、種々の外れ値検出手法やこれまで行われたバースト検出手法を比較することにより、3 σ 法によるバースト検出が妥当であると判断した。

また、どのようなイベントによってバーストが生起するのかについては、様々な要因によってバーストが生起するが、特に地震などの災害によってバーストが生起しやすいことや、他のメディア、とりわけ速報性の高いメディアと関係があることが明らかになった。また、バーストは、「怖」、「哀」などのネガティブな感情と関連するものが全体の 30% 以上とバーストの多くを占めていること、バーストした際にツイート数が多くなるのは、ポジティブな感情が見られるバーストであることが分かった。さらに、地震バーストに影響を与える要因の算出により、最大震度と都心から被災地までの距離の短さがバースト生起の有無に影響を及ぼすこと、特に都心から被災地までの距離の短さがバースト生起により強い影響を及ぼすことが明らかとなった。

最後に、バースト時のツイートの特徴としては、バースト時と非バースト時の比較により、ツイートの文字数が短く、リツイート（RT）の比率が高く、リプライ（@）の比率が低いといった傾向が明らかとなった。このことからバースト時には特に情報入手、拡散ツールとしての側面が強くなることが示された。また、バーストの類型化を行い、各クラスタごとにイベントの特徴を推測することで、各々のバーストは、「小さなイベント型」、「既知イベントピーク型」、「既知イベント準備期間型」、「突発的イベント型」、「情報拡散型」の 5 つに分類できることが分かった。

以上の分析により、Twitter 上でのユーザの情報行動、および Twitter のメディア特性の一端を表すことができたと考えられる。

今回は、バーストについて巨視的観点から分析を行った。しかし、今後の展望としては、より詳細な分析を行うために、ユーザに着目した分析や、特定のバーストを対象として分析を行う必要がある。

謝辞

本修士論文は、筆者が筑波大学大学院図書館情報メディア研究科博士前期課程において行った研究を纏めたものである。主指導教員の本学図書館情報メディア系の池内淳准教授には、研究に関して終始ご指導、ご助言をいただきましたこと心より感謝致します。学群時はクラス担任として、研究室配属後は指導教員として、6 年間にわたり、親身にご指導を賜りました。本当にありがとうございました。

また、本論文を御精読いただき有用なご助言をいただきました、宇陀則彦准教授、辻慶太准教授に深く感謝致します。加えて、本研究を遂行するにあたっては、佐藤哲司教授、及び、佐藤研究室のゼミ生の方々から、実験環境の整備を始めとしたご支援、及び、分析に関するご指導、ご助言を賜りました。厚く御礼申し上げます。

最後になりますが、ゼミを通じて多くの示唆をいただいた池内研究室の皆様に感謝致します。

引用文献

[1] 前島和弘. ソーシャルメディアが変える選挙: アメリカの事例から. アド・スタディーズ, 2010, vol. 34. http://www.yhmf.jp/pdf/activity/adstudies/vol_34_01_05.pdf, (参照 2014-2-14).

[2] “PRESS RELEASE ユーザーローカル、各政党のソーシャルメディア活用度調査を発表。ネット選挙解禁で政党の LINE 利用が活発化”. User Local. 2013-07-18. http://www.userlocal.jp/news/201307181/, (参照 2013-12-29).

[3] “2013 年 SNS 利用動向に関する調査”. ICT 総研. 2013-05-30. http://www.ictr.co.jp/report/20130530000039.html, (参照 2013-12-29).

[4] Semiocast. “Twitter reaches half a billion accounts More than 140 millions in the U.S.”. Semiocast. http://semiocast.com/publications/2012_07_30_Twitter_reaches _half_a_billion _accounts_140m_in_the_US, (accessed 2013-08-01).

[5] Wickre, Karen. “Celebrating #Twitter7”. The Official Twitter Blog. 2013-03-21. https://blog.twitter.com/2013/celebrating-twitter7, (accessed 2013-08-01).

[6] インターネット白書 2012. インプレスジャパン, 2012, p. 28.

[7] Merriam-Webster, Inc. “social media”. Merriam-Webster Online, http://www.merriam-webster.com/dictionary/socialmedia, (accessed 2013-12-29).

[8] 株式会社インセプト. “ソーシャルメディア 【 social media 】”. IT 用語辞典 e-Words. http://e-words.jp/w/SNS.html, (参照 2013-12-31).

[9] “Facebook Reports Second Quarter 2013 Results”. Facebook Investor Relations. 2013-06-24. http://investor.fb.com/releasedetail.cfm?ReleaseID=780093, (accessed 2013-12-29).

[10] “Form S-1 - Securities and Exchange Commission; Twitter, Inc.”. U.S. Securities and Exchange Commission | Homepage. http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm#toc564001_18, (accessed 2013-12-29).

[11] Winkler, Rolfe. “At 300 Million Users, Google+ Usage Questions Remain”. The Wall Street Journal. 2013-10-26. http://blogs.wsj.com/digits/2013/10/29/at-300-million-users-google-usage-questions-remain/, (accessed 2013-12-29).

[12] “プレスリリース [LINE] LINE、登録ユーザー数が世界 3 億人を突破”. LINE Corporation. 2013-11-25. http://linecorp.com/press/2013/1125636, (参照 2013-12-29).

[13] “2013 年度 第 2 四半期決算説明会資料”. http://v3.eir-parts.net/EIRNavi/ DocumentNavigator/ENavigatorBody.aspx?cat=ir_material&sid=25283&code=2432&ln=ja&tlang=ja&tcat=ir_material&disp=simple&groupsid=9438, (参照 2013-12-29).

[14] “2014 年 6 月期第 1 四半期決算説明会資料”. http://v3.eir-parts.net/EIR/View.aspxs?dcat=tdnet&sid=1107558, (参照 2013-12-29).

[15] “プレスリリース 「Ameba」会員数 3,000 万人突破記念 感謝を込めて「3,000 万会員感謝祭」を実施！ スマートフォン向けにブログ機能を強化”. 株式会社サイバーエージェント, 2013-8-22. http://www.cyberagent.co.jp/news/press/detail/id=7925&season=2013&category=ameba, (参照 2013-12-29).

[16] “2013 年度第 2 四半期 決算説明会資料”. http://v4.eir-parts.net/v4Contents/View.aspx?template=ir_material&sid=25334&code=2121, (参照 2013-12-29).

[17] 谷口真嗣. 短大生のインターネット事情について: 常葉学園短期大学を例に. 常葉学園短期大学紀要. 2011, no. 42, p. 113-120.

[18] Oda Masaomi. The Characteristics of the use of Twitter by Beginners: Study of the applicability to the e-healthcare. IEEE International Conference on Systems, Man and Cyvernetrics. 2011, p. 1268-1273.

[19] Prier, Kyle W.; Smith, Matthew S.; Giraud-Carrier, Christophe; Hanson, Carl L. Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic. SBP’11 Proceedings of the 4th international conference on Social computing, behavioral-cultural modeling and prediction. 2011, p. 18-25.

[20] Teranishi, Yuuichi; Shimojo, Shinji. MONAC: SNS message dissemination over smartphone-based DTN and cloud. Peer-to-Peer Computing 2011. 2011, p. 158-159.

[21] 株式会社インセプト. “SNS 【 Social Networking Service 】 ソーシャルネットワーキングサービス”. IT 用語辞典 e-Words. http://e-words.jp/w/E382BDE383BCE382B7E383A3E383ABE383A1E38387E382A3E382A2.html, (参照 2013-12-31).

[22] Rafe, Needleman. “Twitter’s not a social network?”. CNET, 2011-09-23. http://news.cnet.com/8301-19882_3-20112261-250/twitters-not-a-social-network/, (accessed 2013-12-31).

[23] Douglas, Nick. “Twitter blows up at SXSW Conference”. Gawker. 2007-12-03. http://gawker.com/243634/twitter-blows-up-at-sxsw-conference, (accessed 2013-12-29).

[24] “Development of Twitter Services”. MediaOnTwitter. 2012-08-01. http://www.mediaontwitter.com/99/development-of-twitter-services, (accessed 2013-12-30).

[25] Twitter, Inc. ヘルプセンター. https://support.twitter.com/, (参照 2013-12-29).

[26] ツイナビ | ツイッター (Twitter) の使い方. http://twinavi.jp/guide/section/twitter/glossary/%E3%83%AA%E3%83%97%E3%83%A9%E3%82%A4%EF%BC%88%E8%BF%94%E4%BF%A1%E3%83%BB@%EF%BC%89%E3%81%A8%E3%81%AF, (参照 2013-12-31).

[27] Java, Akshay; Song, Xiaodan; Finin, Tim; Tseng, Belle. Why We Twitter: Understanding Microblogging Usage and Communities. In Procededings of the Joint 9th WEBKDD and

1st SNA-KDD Workshop 2007. 2007, p. 56-65.

[28] Krishnamurthy, Balachander; Gill, Phillipa; Arlitt, Martin. A Few Chirps About Twitter. In Proceedings of the First Workshop on Online Social Networks. 2008, p. 19-24.

[29] Kwak, Haewoon; Lee, Changhyun; Park, Hosung; Moon, Sue. What is Twitter, A Social Network or a News Media?. In Proceedings of the 19th International Conference on World Wide Web. 2010, p. 591-600.

[30] Wu, Shaome; Hofman, Jake M.; Mason, Winter A.; Watts, Duncan J. Who Says What to Whom on Twitter. In Proceedings of the 20th International Conference on World Wide Web. 2011, p. 705-714.

[31] Poblete, Barbara; Garcia, Ruth; Mendoza, Marcelo; Jaimes, Alejandro. Do All Birds Tweet the Same? Characterizing Twitter Around the World. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011, p. 1025-1030.

[32] Kivran-Swaine, Funda; Govindan, Priya; Naaman, Mor. The Impact of Network Structure on Breaking Ties in Online Social Networks: Unfollowing on Twitter. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011, p. 1101-1104.

[33] Macskassy, Sofus A.; Michelson Matthew. Why Do People Retweet? Anti-Homophily Wins the Day!. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. 2011, p. 209-216.

[34] Mendoza, Marcelo; Marcelo, Barbara; Castillo, Carlos. Twitter Under Crisis: Can we trust what we RT?. Proceedings of the First Workshop on Social Media Analytics. 2010, p. 71-79.

[35] Suh, Bongwon; Hong, Lichan; Pirolli, Peter; Chi, Ed H. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. Proceeding SOCIALCOM '10 Proceedings of the 2010 IEEE Second International Conference on Social Computing. 2010, p. 177-184.

[36] Ye, Shaozhi; Wu, Felix. Measuring Message Propagation and Social Influence on Twitter.com. Proceedings of the Second international conference on Social informatics. 2010, p. 216-231.

[37] Galuba, Wojciech; Aberer, Karl; Chakraborty, Dipanjan; Despotovic, Zoran; Kellerer, Wolfgang. Outtweeting the Twitterers: Predicting Information Cascades in Microblogs. Proceedings of the 3rd conference on Online social networks. 2010, p. 1-9.

[38] Paul, Sharoda A.; Hong, Lichan; Chi, H. Is Twitter a Good Place for Asking Questions? A Characterization Study. International AAAI Conference on Weblogs and Social Media.

2011.

[39] Antoniades, Demetris; Athanasopoulos, Elias; Polakis, Iasonas; Ioannidis, Sotiris; Karagiannis, Thomas; Kontaxis, Georgios; Markatos, Evangelos P. we.b: The web of short URLs. Proceedings of the 20th international conference on World Wide Web. 2011, p. 715-724.

[40] Abel, Fabian; Gao, Qi; Houben, Geert-Jan; Tao, Ke. Analyzing Temporal Dynamics in Twitter Proles for Personalized Recommendations in the Social Web. Proceedings of the ACM WebSci'11. 2011.

[41] Diao, Qiming; Jiang, Jing; Zhu, Feida; Lim, Ee-Peng. Finding Bursty Topics from Microblogs. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012. p.536-544.

[42] 白木原渉, 大石哲也, 長谷川隆三, 藤田博, 越村三幸. Twitter における流行語先取り発言者の検出システムの開発. 情報処理学会研究報告, データベース・システム研究会報告. 2010, 2010-DBS-150, no. 2, p. 1-8.

[43] Kleinberg, Jon. Bursty and hierarchical structure in streams. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002, p. 91-101.

[44] ITmedia Inc.“「NEC ビッグローブ」による調査データ一覧 | 調査のチカラ”. ITmedia. 2014-1-10. http://chosa.itmedia.co.jp/providers/NEC%E3%83%93%E3%83%83%E3%82%B0%E3%83%AD%E3%83%BC%E3%83%96, (参照 2014-01-13).

[45] “FrontPage/Project311/trend analysis”. Laboratory of Inui and Okazaki. 2012-10-13. http://www.cl.ecei.tohoku.ac.jp/index.php?Project%%E3%83%88%E3%83%AC%E3%83%B3%E3%83%89%E5%88%86%E6%9E%9, (参照 2012-09-30).

[46] Grubbs, Frank E. Procedures for Detecting Outlying Observations in Samples. Technometrics. 1969, vol. 11, no. 1, p. 1-21.

[47] Hawkins, Douglas M. Identification of Outliers. Chapman and Hall. 1980, 188p.

[48] Barnett, Vic; Lewis, Toby. Outliers in statistical data. 1994, 584p.

[49] 小林克己, 金森雅夫, 大堀兼男, 竹内宏一. げっ歯類を用いた毒性試験から得られる定量値に対する新決定樹による統計処理の提案. 産衛誌. 2000, no. 42, p. 125-129.

[50] Dean, Robert B.; Dixon, William j. Simplified Statistics for Small Numbers of Observations. Analytical Chemistry. 1951, vol. 23, no. 4. p. 636-638.

[51] 関哲朗, 後藤勝, 横山真一郎. 正規母集団における順序統計量にもとづく複数外れ値の検出法. 日本経営工学会誌. 1994, vol. 44, no. 6, p. 535-540.

[52] Kadota, Koji; Ye, Jiazhen. Nakai, Yuji; Terada, Tohru; Shimizu, Kentaro. ROKU: An

Improved Method for the Detection of Tissue-Specific Expression Patterns. BMC Bioinformatics. 2006.

[53] 門田幸二. “解析 | 発動変動 | 多群間 |ROKU (Kadota_2006)”. (R で) マイクロアレイデータ解析. http://www.iu.a.u-tokyo.ac.jp/ kadota/r.html#selective_entropy_ROKU, (参照 2013-10-11).

[54] “TCC: Differential expression analysis for tag count data with robust normalization strategies”. Bioconductor. http://bioconductor.org/packages/release/bioc/html/TCC.html, (参照 2013-10-11).

[55] 石川栄助. 棄却検定の比較表. 岩手大学学芸学部研究年報. 1960, vol. 15, no. 2, p. 1-7.

[56] Sprent, Peter; Smeeton, Niqel A. Applied Nonparametric Statistical Methods, Chapman and Hall. 1993. 480p.

[57] Sprent, P. Data Driven Statistical Methods, Chapman and Hall. 1997. 406p.

[58] Maronna, Ricardo A.; Martin, Douglas R.; Yohai, Victor J. Robust Statistics: Theory and Methods, Wiley. 2006. 436p., (Wiley Series in Probability and Statistics).

[59] Romero, Daniel M.; Meeder, Brendan; Kleinberg, Jon. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In Proceedings of the 20th International Conference on World Wide Web. 2011, p.695-704,

[60] Dan, Chalmers; Simon, Fleming; Iwan Wakeman; Des, Watson. Rhythms in Twitter. 2011 ieee third international conference on social computing (socialcom). 2011, p. 1409-1414.

[61] 榊剛史, 松尾豊. ソーシャルセンサとしての Twitter: ソーシャルセンサは物理センサを凌駕するか?. 人工知能学会誌. 2012, vol. 27, no. 1, p. 64-74.

[62] Bollen, Johan; Pepe, Alberto; Mao,Huina. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. 2011, p.450-453.

[63] Asur, Sitaram; Huberman Bernardo A. Predicting the Future With Social Media. In Proceeding WI-IAT'10 Proceedings of the 2010 IEEEWICACM International Conference on Web Intelligence and Intelligent Agent Technology. 2010, p. 492-499.

[64] Tumasjan, Andranik; Sprenger, Timm O. Sandner, Philipp G. Welpe, Isabell M. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 2010, p. 178-185.

[65] Takeshi, Sakaki; Makoto, Okazaki; Yutaka, Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors, In Proceedings of International Conference

on World Wide Web. 2010, p. 851-860.

[66] NEC ビッグローブ. “ついっぷるトレンドとは?”, http://tr.twipple.jp/about/, (参照 2013 年 10 月 20 日)

[67] Dodds, Peter Sheridan; Harris, Kameron Decker; Kloumann, Isabel M.; Bliss, Catherine A.;Danforth, Christopher M. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. PLoS ONE. 2011, vol. 6, no. 12.

[68] 三浦麻子. 東日本大震災とオンラインコミュニケーションの社会心理学: そのときツイッターで何が起こったか. 電子情報通信学会誌. 2012, vol. 95, no. 3, p. 219-223.

[69] 中村明. 感情表現辞典. 東京堂出版, 1993, 458p.

[70] 原田実. 意味解析に基づくテキストマイニングシステム STM. 電子情報通信学会技術研究報告 NCL, 言語理解とコミュニケーション. 2011, vol. 110, no. 400. p. 29-34.

[71] 松本和幸, Bracewell, David B., 任福継. 黒岩眞吾. 感情コーパス作成支援システムの開発. 自然言語処理研究会報告. 2005, no. 117. p. 91-96.

[72] 曽我幸雅, 中村岳史, 山田達也, 濱川礼. 発言者の感情を取得しグラフィカルに表現するシステム. 全国大会講演論文集. 2008, 第 70 回, no. 4, p. 235-236.

[73] 東樹和美, 古川敦子. 感情表現を中心とした学習項目の提案に向けて. 日本語教育方法研究会誌. 2003, vol. 10, no. 2, p. 8-9.

[74] 齊藤崇子, 中村知靖. 日本人における情動概念の階層構遊. 九州大学心理学研究. 2003, vol. 4, p. 95-99.

[75] 相馬壽明. 関根弘子. 聴覚障害児童・生徒の語彙に関する研究: 感情語を用いて. 特殊教育学研究. 1986, vol. 24, no. 2, p. 27-34.

[76] 原田隆史. 江藤正己. 高柳知世. 書評を用いた図書に対する感性パラメータの自動設定. 情報知識学会誌. 2008, vol. 18, no. 2, p. 153-160.

[77] 青木翔, 内田理. ブログを用いた絵文字の感情ベクトル作成手法. 電子情報通信学会技術研究報告, NCL, 言語理解とコミュニケーション. 2011, vol. 110, no. 400. p. 25-28.

[78] 橋本和幸, 中川博之, 田原康之, 大須賀昭彦. 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. 電子情報通信学会論文誌, D, 情報・システム. 2011, vol. J94-D, no. 11. p. 1762-1772.

業績一覧

学会発表

[1] 水沼友宏, 池内淳. 学会ウェブサイトのアクセスログ分析: 日本図書館情報学会ウェブサイトを対象として. 日本図書館情報学会, 2012 年 5 月 5 日.
[2] 水沼友宏, 山口裕太郎, 山本修平, 島田論, 池内淳, 佐藤哲司. Twitter におけるバースト状態に関する実証的研究. 情報社会学会第 5 回知識共有コミュニティワークショップ. 2012 年 11 月 10 日. (論文投稿推薦)
[3] 水沼友宏, 菅原真紀, 池内淳. 大学生の Twitter における行動規範に関する分析. 情報社会学会年次研究発表大会. 2013 年 5 月 25 日. （プレゼンテーション賞受賞）

査読付き原著論文

[4] 水沼友宏, 山口裕太郎, 山本修平, 島田論, 池内淳, 佐藤哲司. Twitter におけるバーストの生起要因と類型化に関する分析. 情報社会学会誌, Vol.7, No. 2, 2013, p. 23-38.
[5] 水沼友宏, 菅原真紀, 池内淳. 大学生の Twitter における行動規範に関する分析. 情報社会学会誌, Vol. 8, No.1, p. 23-38