卒業論文　　2012年度（平成24年度）

Twitterを用いた状況検知システムの設計と構築

慶應義塾大学 環境情報学部

倉田 彩子

卒業論文要旨 - 2012 年度 (平成 24 年度)

Twitter を用いた
状況検知システムの設計と構築

社会の情報化、通信デバイスの普及によって、24 時間いつでもどこでも情報の入手・発信が可能となった。これに伴い、人々の興味・関心の対象が変化するスピードも速くなっている。さまざまなイベントへの参加や購買活動など、私たちの行動は流行に左右されることが多く、各種メディアは、様々な方法で“今”話題となっているものをいかに早く発信するか試行錯誤している。

そこで、本研究では Web 上に存在するテキストや位置情報を用いることで、“今”この瞬間に起こっている話題のものとそれに付随する場所を検知する手法を提案する。

本論文では、マイクロブログサービスの 1 つである Twitter に着目し、Twitter に投稿されるテキスト/位置情報と実空間で発生するイベントとの間の相関について調査・検証を行った。その結果、Twitter に投稿されるテキストの数、テキスト内容、位置情報は、実空間イベントから大きな影響を受けていることがわかった。このことから、Twitter への投稿の収集、解析を行うことで実空間の状況検知が可能であるという仮説を立て、これを実現するシステム、AKT24(Ayako Kurata Tweet-analyzer 24h) の設計と構築を行った。

AKT24 は、Twitter に投稿された情報を、字句/時間/緯度/経度の 4 次元で解析を行うことで実空間の状態検知を行うことを目的とする。システムは地図と、横軸をキーワード、縦軸を時刻とするグラフで構成され、これによって指定期間中の特徴キーワード、それを含むツイートの地理分布を表示する。キーワードの抽出には、通常時との出現数の差分の大小を利用する手法を採用した。

システムの実装、評価の結果、本システムにより複数の実空間イベントの発生とその発生箇所が感知できることが確認された。

本研究によって、これまで情報収集のために必要とされていた時間やコストを大幅に削減することが可能となる。

キーワード:
1. Twitter, 2. 時空間解析, 3. 位置情報

慶應義塾大学 環境情報学部

倉田 彩子

Abstract of Bachelor's Thesis - Academic Year 2012

Auto situation detecting system using Twitter

Today, the computerization and the popularization of communication devices have made it possible to get and provide the information at anytime, anywhere.

Accordingly, the things and places we become interested in change quickly and what event we should take part in or what we should buy can be easily influenced by the fashion. Thus, various media struggle how to detect and provide information about “ hot topics ” .

We have investigated and validated the correlation between texts, geo-information of Twitter and events in real world. As a result, we found that the number, contents of texts and geo-information are influenced by events in real world. From above, we made a hypothesis which is “ we can detect situation in real world by collecting and analyzing tweets ” and built the system, “ AKT24(Ayako Kurata Tweet-analyzer 24h) ” to validate the hypothesis.

The purpose of this system is to detect present situation in real world by analyzing tweets from the point of text, time, latitude and longitude. This system consists of map and graph, showing characteristic keywords on longitudinal axis and time on horizontal axis. This indicates characteristic keywords and its geographical distribution in specific period. To extract the keywords, we use the difference of frequency of words used between specific period and previous one.

After this implementation and validation of this system, we verified that we were able to detect what and where the events occur in real world.

This study contributed the dramatic reduction in search costs and time for “hot topics ” .

Keywords :
1. Twitter, 2.Temporal-spatial analysis, 3.Geo-information

Keio University, Faculty of Environmental information

Ayako Kurata

目 次

図 目 次

表 目 次

1　序論

1.1　はじめに

社会の情報化が進み、様々な情報を誰もが簡単に手に入れることができるようになった。これに伴い、人の趣味や嗜好は多種多様となり、興味関心が移り変わるスピードも早くなった。この急速な変化に対応するように、様々な形態のメディアや店舗が登場している。その１つとして、新商品や毎日変化する売れ筋商品のみをランキング形式で販売する形態の店舗 [1] がある（図 1）。この店舗で陳列される商品は流通各社のデータをもとに決定され、約 2 週間で移り変わる。この販売形態は、これまでの、店や商品が流行を創り出すという従来のスタイルを、流行が店を作るという新しいものへ変えたといえる。

このように、販売の現場では流行を察知する、または創り出すということが非常に重視され、その方法は日々試行錯誤されている。

流行を発信する代表的なメディアとして雑誌やテレビがある。これらの情報は多くの調査に基づいて発信されるため、一般に正確性が高いとされているが、それゆえに人々の手に渡るまでの時間やコストがかかる。

また今日では、インターネット上でも企業・個人ブログやクチコミサイトといった様々な形でトレンド情報が発信されている。これらのサイトでは、商品や飲食店を利用した客がその商品についてレポートし評価をつけることで、数ある商品のランク付けを行う。これから商品を購入しようとしているユーザは、これらの情報を商品選択の参考とする。これらの手段は、情報が受け手に辿りつくまでの時間やコストが雑誌やテレビに比べて大幅に削減される一方で、ユーザが意識的にクチコミ等を提供する必要があるため手間がかかり、また情報の一般性、正確性の欠如という問題がある。そして、これらのサイトを成立させるのは、レポートや評価を提供するユーザの善意であるというもろさが存在する。

出所）流行発信ショップ「ranKing ranQueen」, http://www.ranking-ranqueen.net/pressrelease/pdf/20130116_2.pdf

図 1: ranKing ranQueen 概要

流行の急速な変化を創り出している要因の１つにSNS(Social Networking Service)やマイクロブログサービスの普及が挙げられる。これらのサービスは、１つ１つの小さな情報を瞬時に拡散させることを可能にした。これにより、これまでは注目されることのなかった単なるつぶやきが、共感され拡散されることで大きな力を持つようになった。

SNSやマイクロブログサービスの利用者は今後も増え、Web上を行き交う声も増え続けると考えられる。ウェブから生まれる流行も刻々と変化していく。

そのような状況の中では、流行をいち早くキャッチし、人々の行動決定に役立てる手段が求められる。そこで、このWeb上を行き交う大量の声を分析し、意味を読み取ることで、ユーザの手間なく、十分に一般性を持った世の中の流行を検知する手法を提案する。

本研究では、マイクロブログサービスのTwitter上に発信されるテキストと、それに付与された位置情報を用いたイベント検知システムの設計と実装を行う。

1.2 目的

個人によって発信されるテキスト情報および位置情報を用いて、社会の流行やホットトピックを検知することを目的とする。そのために、大量のテキスト情報および位置情報の管理、解析、視覚化を行うシステムの設計と構築を行う。

1.3 本論文の構成

本論文は全８章で構成する。

第２章で、背景となるサービスとその利用状況、またその社会的影響を示す。第３章で、本研究で利用するマイクロブログサービスであるTwitterについて、その実空間との相関を事例と独自に行った検証の結果を用いて示す。

第４章では、関連する研究/サービスの紹介と比較を行い、本研究が目指す姿を示す。第５章、第６章で、本研究で構築したシステムについての設計と実装について述べ、第７章で、構築したシステムの評価を行う。第８章で、本研究から導かれた結論と今後の課題を示し、まとめとする。

2　背景

本章では、本研究の背景であるソーシャルメディア、位置情報サービス、集合知の現状について述べる。

2.1　ソーシャルメディア

本論文では、ソーシャルネットワーキングサービス (SNS) とマイクロブログサービスをまとめたものをソーシャルメディアとする。両サービスについての説明を示す。

SNS は、社会的なつながりをインターネット上でも実現させるサービスである。代表的な SNS として、Facebook[2] や mixi[3]、LinkedIn[4] といったものが存在する。SNS の中には実名を必須とするもの/しないもの、日記や写真をメインとしたものやゲームを主体としたもの、ターゲットをビジネスの現場に絞ったもの等、さまざまな種類がある。

マイクロブログサービスは、200 文字程度の短い文章を投稿するブログサービスである。代表的なものとして Twitter[5] がある。近年さまざまな場面で利用され、注目されるサービスである。

ソーシャルメディアには、各社が提供する様々なサービスがある。資料 [6] によると、ソーシャルメディア利用者の利用目的も「リアルな友人とのコミュニケーション」、「暇つぶし」、「ネット上の知り合いとのコミュニケーション」、「趣味などに関する情報収集」などと様々である。ソーシャルメディア同士や他のサービスとの連携も進んでいることから、インターネット上の様々なサービスのプラットフォームとしても注目を集めている。また、ソーシャルメディアの利用人口は年々増加している。特に近年の増加幅は大きく、2012 年 5 月時点での日本国内のソーシャルメディア人口の推定値は 5060 万人と、2011 年の同人口に比べ、1530 万人の増加がみられた。2008 年からの同人口の推移を図 2 に示す。

また、図 3 から、スマートフォン利用者は、他の端末利用者に比べ、mixi、Facebook、Twitter といったソーシャルメディアの利用者が多いこともわかる。ソーシャルメディアは、固定された室内でなく、外出先や移動中からも頻繁に利用されていると考えられる。

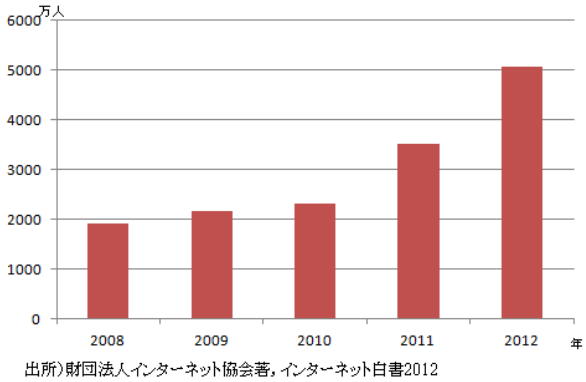


図 2: ソーシャルメディア人口の推移

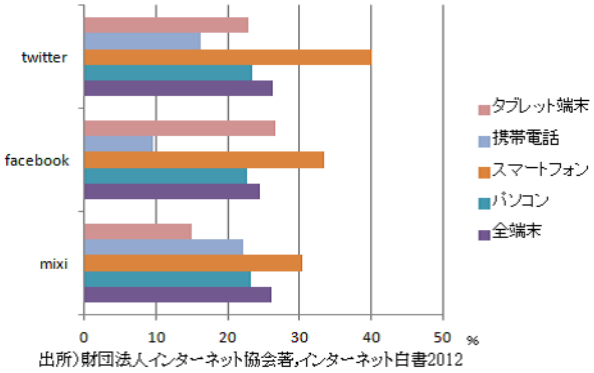


図 3: 利用しているソーシャルメディア

2.2 Twitter

本節では、ソーシャルメディアの中でも近年特に注目を集めている Twitter について詳説する。

Twitter とは、米 Obvious 社（現 Twitter 社）が 2006 年 7 月に開始したマイクロブログサービスである。ユーザは、1 度に 140 文字以内で文章を投稿することができる。この文章のことをツイートと呼び、投稿することを「つぶやく」、「ツイートする」という。ログインすると、タイムラインと呼ばれる画面が表示され、他のユーザが投稿したツイートを時系列に読むことができる。Twitter ユーザの間には「フォロー」「フォロワー」という関係があり、興味のあるユーザを「フォロー」することで自分のタイムラインにそのユーザのツイートが表示されるようになる。反対に自分のことをフォローしたユーザを「フォロワー」と呼ぶ。ツイートは非公開にすることもでき、自分がフォローされる際に、フォロワーの許可/不許可を選択することができる。ツイッターのメイン画面例を図 4 に示す。

図 4: Twitter メイン画面例

投稿機能には、ツイートの他に、リツイート、リプライがある。

リツイートとは、他のユーザが投稿したツイートを再投稿することで、この機能は Twitter の大きな特徴の 1 つとされる。リツイートには、元のツイートをそのままの形で投稿する公式リツイートと、元のツイートを引用し、自分のコメント等を付け足して投稿する非公式リツイートがある。一般に、リツイートされた回数が多いほど、そのツイートの注目度・重要度は高いとされ、現在はリツイートされた回数が多かったツイートを知らせるサービスが複数存在する [7],[8]。

リプライとは、「@宛先アカウント名」をツイートに記述することで、特定のユーザに向けてツイートすることである。この投稿は、宛先、送り主、宛先と送り主双方をフォローしているユーザのタイムラインにのみ表示され、チャットのように使われることもある。

投稿機能の他、ハッシュタグも Twitter の特徴の一つである。ハッシュタグとは、同じ話題のツイートに付与される目印のようなものである。「#ハッシュタグ名」をツイートに記述するだけで、ツイートをグループ化したり、同じ話題のツイートを検索してまとめて読むことができる。現在、ハッシュタグ分析・検索サービス [9] も存在する。

また、ツイートにはジオタグと呼ばれる位置情報を付与することができる。この機能により、ツイートのテキスト情報とともに緯度・経度情報が投稿され、地図でユーザの現在位置が示される。ジオタグが付与されたツイートの例を図 5、図 6 に示す。

Twitter は 21ヶ国語に対応しており、世界中で利用されている。2012 年現在、全世界でのアカウント数は 4 億 6500 万を超えており、1 日に 1 億 7500 万のツイートが投稿されている [10]。また、国別アカウント数は、1 位がアメリカで 1 億 770 万アカウント、2 位はブラジルの 3330 万アカウント、3 位に日本の 2990 万アカウント [10] と、世界的に見て日本の Twitter 利用者は多い。図 7 に国別アカウント数を示す。

図 5: ジオタグ付ツイート例

図 6: 位置情報表示例

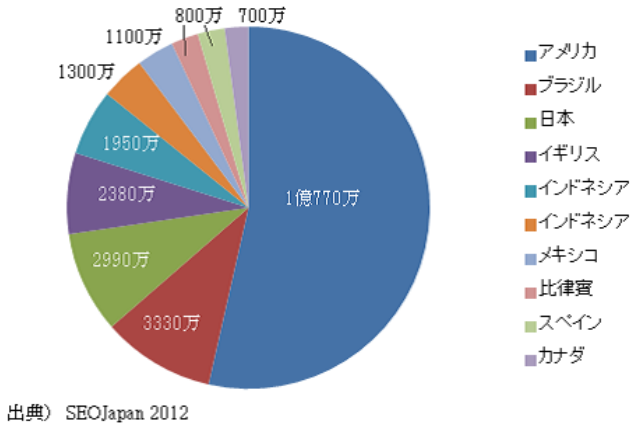


図 7: 国別アカウント数

2.3 位置情報

現在、携帯電話や携帯ゲーム機、スマートフォンなど、位置情報の発信を可能とするデバイスは多く存在する。図 8 が示す通り、インターネット利用者数全体

に占めるモバイルデバイス利用者の比率は 2002 年を境に急激に増加している。この流れは、室内に限られていたインターネット利用環境を、外出時などいつでもどこでも利用できる、ユーザの行動スタイルに合ったものへと変化させた。この動きの中で、ユーザが発信する位置情報からその行動パターンを調査する手法といった、ユーザの位置を利用した研究が多く行われている。その 1 つとして、酒巻ら [11] の研究が挙げられる。

酒巻ら [11] は、Twitter に投稿されたテキストと位置情報から、その位置が投稿したユーザにとってどのような意味を持つかという情報を推定する手法を提案した。これを実現させることで、Twitter を用いて人の行動調査を行うことが可能となる。提案手法は、まずツイートの位置情報により、ツイートのクラスタリングを行う。次に、各クラスタ内の投稿内容に形態素解析を行い、そのクラスタを代表する単語を抽出する。提案手法の結果、「起きる」、「寝る」、「家」といった単語のグループが検出され、その範囲が「自宅」に関する箇所であることが推測できた。

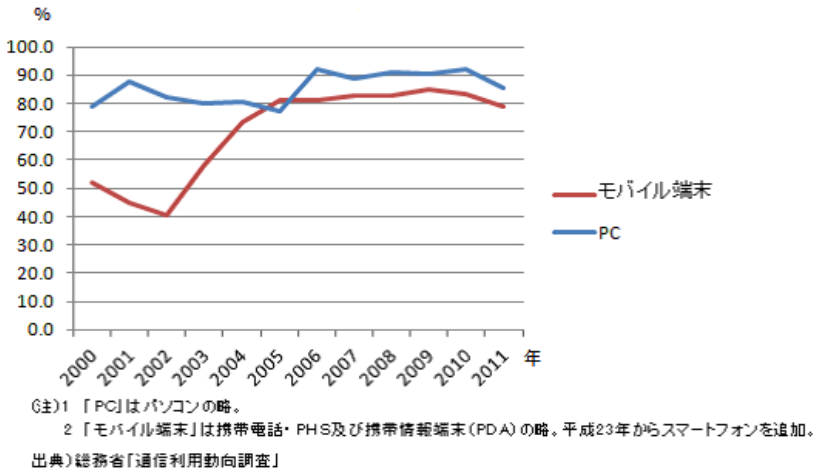


図 8: 端末別に見たインターネット利用者数・比率の推移

モバイルデバイスの普及に伴い、位置情報を利用したサービスも増加している。位置情報を利用したサービスは、おおまかに

- コミュニケーション/ライフログ
- ロケーションベースマーケティング
- ゲーム

の 3 つに分類することができる。

コミュニケーション/ライフログ系の位置情報サービスの代表例として、foursquare[12]、ロケタッチ [13]、Google Latitude[14]、Facebook/Twitter のロケーション機能等が挙げられる。例として、foursquare の概要を示す。

foursquare は、スマートフォンや携帯電話を使って、自分の位置を他のユーザと共有したり、他のユーザへリコメンドすることができるサービスである。例えば、ある飲食店へ行った際、チェックイン機能を使ってメッセージを残したり、過去にどんな人がチェックインしたかを知ることができる。また、チェックインの回数によって店舗からサービスが受けられるなど、店舗と連携したサービスも行っている。Twitter や Facebook と連携して利用することで、自分が訪れた場所を友人と共有するユーザも増えている。このように、コミュニケーション/ライフログ系位置情報サービスは、位置を媒体として人とのつながりを創り出す、SNS の新たな形として注目を集めている。

図 9: foursquare 操作画面

ロケーションベースマーケティング [15] は、近年利用が増加しつつあるマーケティング手法で、企業等がユーザの位置情報と連動して情報配信を行う。例えば、ユーザの現在位置から、周辺のエリア情報を提供したり、目的の商品が買える一番近い店舗を紹介する。位置情報を用いたマーケティングの例としては、日立製作所の地図クル [16] が挙げられる。位置情報を用いたマーケティング手法については、位置情報から得られる購買行動の記録から消費傾向や消費意欲の関係を調査する研究 [17] も行われている。飯尾ら [17] らは、携帯電話から得られる位置情報ログと、オンラインアンケートの結果を結び付けることで、実際の行動と消費意欲の関係を分析した。平日および休日に消費者が過ごす行動についてあらかじめ 4 つのタイプを用意し、得られた位置情報ログから、被験者の行動をこれらのタイプに分類する。検証の結果、「特定の行動タイプの被験者はファッションにお金をかける傾向がある」、「特定の行動タイプの被験者は独身や単身者が多い」など、行動タイプごとに属性や購買行動の特徴がみられた。

位置情報を利用したゲームは、2000 年に登場して以来、ゲーム市場で人気を博している。位置情報ゲームでは、実際の移動距離に応じてポイントを取得し、それを使って仮想的な町を作り上げるもの、実際にある場所へ行って位置情報を送信することでその地点を仮想的に「統一」し、これを繰り返しながら日本中の統一を目指すもの等、様々なものがリリースされている。中でも人気を博しているのが、株式会社コロプラが運営するコロニーな生活☆ PLUS[18] である。その概要

を示す。

ゲームに登録すると、自分だけの街=コロニーが作成される。コロニーな生活は、そのコロニーを発展させていく街育成ゲームである。育成するためにはゲーム内の通貨である「プラ」が必要となる。この「プラ」を取得するには実際に移動しなければならず、その移動距離に伴って、取得できる「プラ」も増える。1km なら 1 プラ、10km なら 10 プラ取得できる。また、限られた場所でのみ買うことができるお土産やスタンプもあり、これらのシステムがユーザの移動を促す。このゲームは、2005 年にリリースされて以来ユーザ数を増やし続け、2012 年 7 月の段階でユーザ数は 300 万人を突破した [19]。また、2011 年 6 月には東急百貨店吉祥寺店と連携し、コロプラ物産展 2011[20] が開催された。このイベントは 9 日間の開催で 4 万人を動員し、さらに売上合計は約 7000 万円、この会場規模として開店以来の売上を記録した。遠方からの来場者も非常に多く、「一都三県以外」の来場者だけで通常の週末並みの来客を記録するなど、集客効果を発揮した。

この事例は位置情報ゲームが人の実空間での行動を促した例といえる。

2.4 集合知 (クチコミ)

2.1 節で述べたソーシャルメディアの普及の結果、今日のインターネット上には個人の感情や感覚、感想を記したテキストデータで溢れている。このようなデータは、一般に非構造化データ [21] の一つとされる。

このような非構造化データは、多く集めることで人の行動決定や購買意欲に影響を及ぼし得る有用な情報となる。それを利用したサービスの例が、クチコミサイトや商品のレビューサイトである。

購買者の欲しい商品・サービスの情報源についての調査結果を図 10 に示す。欲しい商品・サービスの情報源としてクチコミサイトや商品などのレビューサイトを挙げた人の割合は、企業のウェブサイトや商品・サービス提供者からのメールマガジン等に比べ高い [10]。このことから、消費者は、各企業の発表する情報よりも実際に利用した個人の感想を重要視していることがわかる。さらに、商品のクチコミを発信するサイトでも、「各専門分野の商品・サービスを紹介する紹介サイト（個人が運営するもの)」や「個人ホームページ」の順位が低いことから、インターネット上のクチコミは、大量に集まった状況で有用となると考えられる。

2.5 本章のまとめ

本章では、ソーシャルメディア、位置情報サービス、集合知（ロコミ）を利用したサービスの種類・普及状況と実社会への影響について示した。

ソーシャルメディアの利用者は年々増加しており、その利用目的は友人とのコミュニケーションや暇つぶしなど、複数あることがわかった。

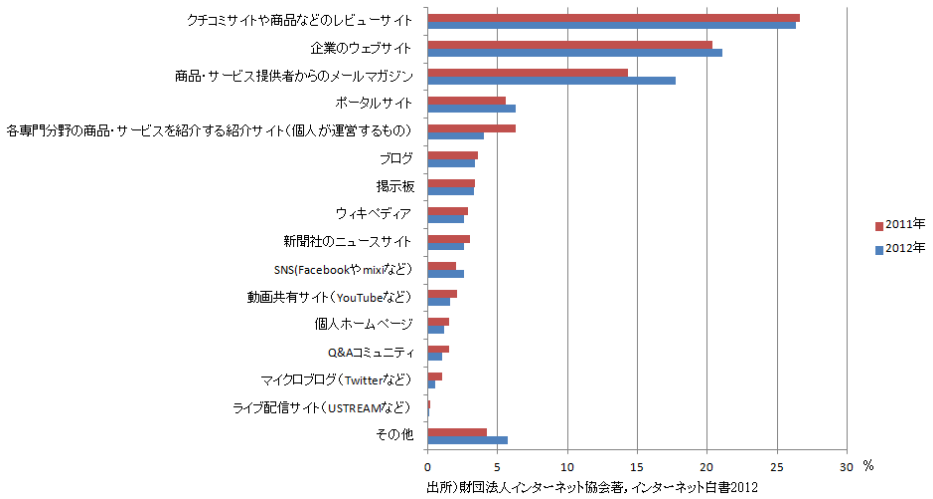


図 10: 欲しい商品・サービスの情報源

位置情報サービスについては、その利用について、コミュニケーション/ライフログ、ロケーションベースマーケティング、ゲームの大まかに 3 つのジャンルに分類される。特にゲームの分野では、提携した百貨店イベントで大きな売り上げを上げるなど、実空間に多大な影響を及ぼした。

集合知を利用したサービスは、購買者が購買活動をする上で重要視されている。さらに、Web 上に存在するクチコミは、大量に集まった状況で有用となると考えられることがわかった。

以上を踏まえ、次章では対象をマイクロブログサービスである Twitter に絞り、Twitter と実空間イベントとの相関について、複数の事例と独自に行った検証を用いて示す。

3　実空間と Twitter

本章では、2 つの事例と独自に行った事前検証の結果を用いて、Twitter と実空間イベントとの相関について示す。

本章では、マイクロブログサービス Twitter の概要と、これまでの Twitter と実空間イベントとの連携例を示す。

3.1　「天空の城ラピュタ」テレビ放映時における Twitter

資料 [10] によると、2012 年 12 月 22 日現在の瞬間最高ツイート数トップ 3 は、1 位が映画「天空の城ラピュタ [22]」テレビ放映時（日本、25088 ツイート/秒）、2 位、「スーパーボウル XL」優勝決時点（アメリカ、12233 ツイート/秒、3 位「スーパーボウル」マドンナ登場時（アメリカ、10245 ツイート/秒）と、映画やスポーツの特定のシーンと連動している。

1 位となった「天空の城ラピュタ」テレビ放映時には、特に主人公である 2 人が滅びの呪文「バルス」を唱える瞬間に瞬間最高ツイート数を達成した。この際のツイート内容はほとんどが「バルス」を含むものである [23]。映画を見ていた視聴者が、映画中の特定のシーンと同時にツイートをしたためと考えられる。これは、実空間での出来事の盛り上がりが Twitter 上にも反映された例といえる。

この日の全世界におけるツイート数の変化および「バルス」に関するツイート数の変化を図 11、図 12 に示す。

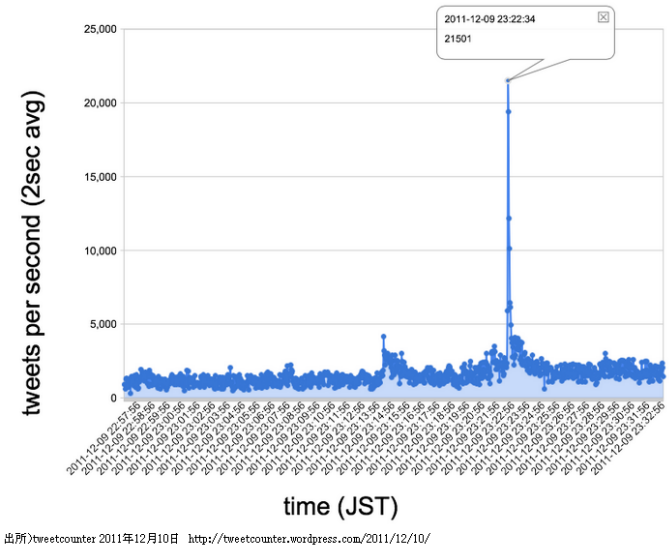


図 11: 天空の城ラピュタ放映時の全世界のツイート数推移

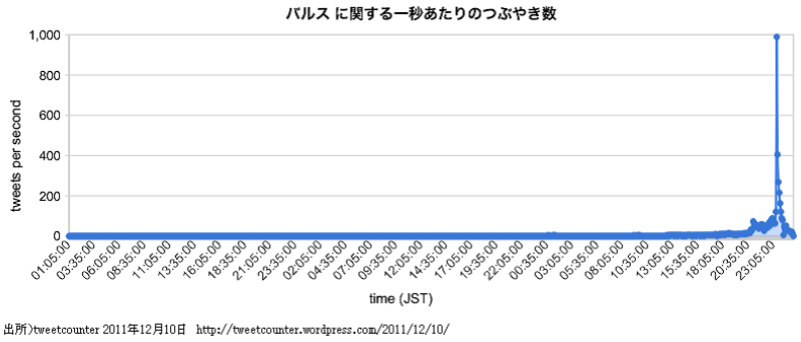


図 12: 天空の城ラピュタ放映時の「バルス」に関するツイート数推移

3.2 東日本大震災発生時における Twitter

2011 年 3 月 11 日に発生した東日本大震災は、震源地である東北地方で多数の犠牲者を出すと同時に、関東地方でも交通期間の乱れや停電、通信手段の不通といった各種の混乱を招いた。そのような状況の中で、Twitter は安否確認や情報発信の手段として大きな役割を果たした。

この際、大きな話題となった取組として「ヤシマ作戦」が挙げられる。「ヤシマ作戦」とはテレビアニメ「新世紀エヴァンゲリオン [24]」に登場する作戦名で、アニメ中では、攻撃兵器の電力を集めるために日本中を停電状態にする作戦を指す。震災時、福島県の原子力発電所をはじめとする発電設備が大きな打撃を受けたため、東北および関東地方で停電の恐れが生じた。この停電を免れるため、「ヤシマ作戦」の実行が Twitter 上で呼びかけられ、日本中で節電の流れが起こった。この出来事は世界的にも Twitter の影響力の大きさが認識されるきっかけとなった。

本研究では、震災発生時におけるツイート内容と時間との関係を調べるため、東日本大震災発生時に関東と東北で投稿されたジオタグ付ツイート数の推移を調査した。結果を図**??**に示す。グラフは、震災発生の前後 1 日ずつ、計 3 日間の推移を示している。

ツイート数のピークは、関東では 11 日 19:30 から 20:30 までの 3017 ツイート、東北では 11 日 15:30 から 16:30 までの 189 ツイート、また 3 日間合計は関東は 52580 ツイート、東北で 3460 ツイートと、ツイート数には約 15 倍の差があった。これは人口の差と Twitter ユーザの差によるものと考えられる。しかし、ツイート数は異なるものの、関東、東北ともに地震が発生した 14:46 を境に急激にツイート数が変化していることがわかる。

次に、表 15 に示すのは地震発生当日 13 時 30 分から翌日 00 時 30 分までの、時間帯ごとの関東でのツイートの頻出単語とその出現数である。11 時間で合計 23980 ツイート取得することができた。この日特に多く見られた単語の出現数の推移を示した。震災当日、Twitter は安否確認や災害情報、交通情報、天気といった様々な情報を得るためのツールとして重要な役目を果たした。頻出単語を時間別に見

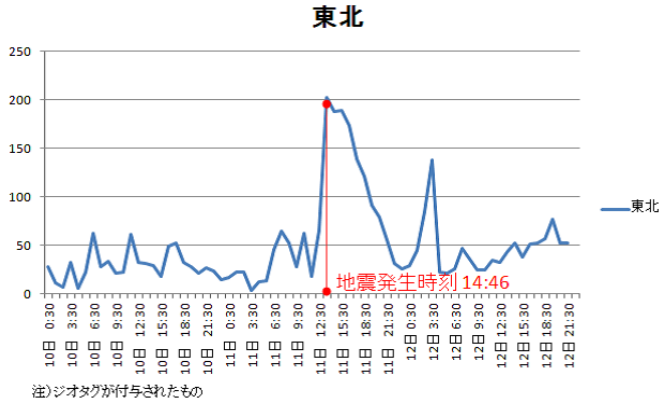


図 13: 東北でのツイート数推移

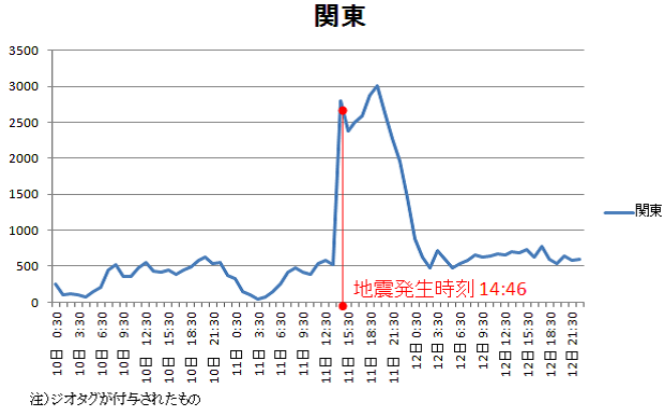


図 14: 関東でのツイート数推移

ても、震災発生前の 13 時 30 分から 14 時 30 分までは「笑」、「人」、「日」といった、それだけで特に意味を表さない単語が多く、またその出現数も少ない。したがって、それぞれのツイートに共通性はあまり見られない。

一方で、震災発生後の 15 時 30 時以降を見てみると、「地震」、「大丈夫」、「余震」、「電車」といったキーワードが多く出現し、またの出現数も格段に増加している。このことから、地震発生をきっかけにユーザが一斉に共通の話題についてツイートしていることがわかる。

これらの単語をその内容からおおまかに以下の 4 つのジャンルに分別し、ジャンル毎に時間に伴う出現数の変化を検証した。

- 安否確認：単語例「大丈夫」「無事」
- 交通：単語例「駅」「線」

- 現在状況の発信：単語例「通過」「帰宅」
- 災害そのものについて：単語例「地震」「揺れ」

その結果、それぞれのジャンルごとに、時間に伴ってその単語の出現数が変化していることがわかった。

災害そのものについてのツイートや安否確認をするようなツイートは、地震発生時をピークに減少傾向にあり、反対に交通網についてのツイートや現在状況の発信をするツイートは時間の経過とともに増加傾向にある。これは、仕事や学校等で中心部へ来ていた人々が、時間が経つにつれて帰宅手段について検討したり、帰宅状況について発信したためと考えられる。

13:30		15:30		16:30		17:30		18:30		19:30		20:30		21:30		22:30		23:30	
5	笑	**41**	**地震**	**34**	**大丈夫**	**34**	**帰宅**	54	人	44	人	39	人	25	家	27	線	**19**	**駅**
5	人	**31**	**大丈夫**	25	無事	26	人	**29**	**帰宅**	**35**	**帰宅**	**35**	**駅**	**22**	**帰宅**	18	家	14	人
3	日	26	無事	**24**	**地震**	26	東京	27	通り	**29**	**駅**	31	線	22	線	**18**	**駅**	14	線
3	今日	22	津波	19	電話	22	バス	**23**	**駅**	27	分	29	帰宅	**20**	**駅**	**17**	**無事**	**12**	**帰宅**
2	どこ	18	避難	**17**	**帰宅**	21	渋滞	23	今日	**27**	**通過**	**22**	**通過**	16	人	**17**	**帰宅**	11	時間
2	雨	16	ビル	16	余震	21	無事	20	渋滞	25	時間	22	ここ	14	区	15	再開	11	今
2	合格	15	みんな	16	駅	21	駅	**20**	**通過**	24	渋滞	21	時間	12	東京	14	時間	11	家
2	今日	15	電話	15	家	**21**	**地震**	19	車	22	家	20	東京	12	休憩	14	渋滞	9	電車
2	丁目	**13**	**帰宅**	15	連絡	18	家	18	東京	20	新宿	19	停電	12	渋滞	12	到着	9	余震
2	都	13	人	15	東京	17	電話	18	バス	19	無事	17	町	11	新宿	11	新宿	8	車

注）2011年3月11日 13:30～23:59の間に関東で投稿されたツイート 全23980ツイート

図 15: 東日本大震災発生時のツイート頻出単語の変化

3.3 事前検証

前節で示した例から、ツイート数、ツイート内容はともに実空間で起こったイベントと密接な関係があると考察した。本研究で構築するシステムの方針を設定するにあたり、この考察についての検証を行う。

3.3.1 検証の概要

実空間イベントと位置情報付きツイートの位置/内容/数は相関するか否かについて事前検証を行う。あらかじめキーワードを設定し、キーワードを含むツイートを、その位置情報から地図上にプロットする。そのキーワードに関するイベントの発生および発生箇所がそこから検知できるか検証する。

キーワードは「花火」に設定し、花火大会が予定されていた 2012 年 7 月 28 日と、予定されていない 2012 年 7 月 27 日でツイートの動向を比較した。時間帯は 16 時から 23 時 59 分で、1 時間ごとに比較を行った。

検証にあたり、次に述べるツイートマッピングシステムを実装した。

3.3.2 ツイートマッピングシステム概要

検証を行うにあたり、独自に実装を行ったツイートマッピングシステムの概要を示す。システムの実装にあたっては、梶原 [25] の研究を参考とした。

本システムは、ツイートの位置を視覚的に認識することを可能とする。投稿された日時や含まれるキーワードを選択することができ、特定の日時、キーワードを含むツイート群について調査することができる。地図とキーワード、対象日時選択欄で構成され、キーワードと日時を選択すると、地図上に該当するツイートがピンのアイコンで表示される。図 16 に操作画面を示す。

なお、検証に用いたツイートは、Twitter 社が提供する streamingAPI を使い、日本列島をカバーする緯度 127.4414〜148.7109 度、経度 29.9930〜45.8900 度の範囲で投稿されたもののみを独自に収集した。検証するにあたって収集したツイートの情報を表 1,2 に示す。

表 1: 収集したツイート情報

user	ユーザ名
date	日時
lat	緯度
lng	経度
text	ツイート内容

表 2: ツイート数

指定期間中の全ツイート	81791
キーワードを含むツイート	1690

図 16: システム画面

3.3.3 検証結果

2011 年 8 月 27 日、2011 年 8 月 28 日で、キーワードを含むツイートの割合、地理分布は明らかに異なった。

花火大会のなかった 27 日は、キーワードを含むツイートの割合は 9 時間平均で 0.56 %、最小が 15 時台の 0.33 %、最大が 20 時台の 0.89 %であった。これに対し花火大会のあった 28 日は、キーワードを含むツイートの割合が 9 時間平均で 3.06 %、最小が 15 時台の 1.17 %、最大が 19 時台の 6.67 %と、最大で 5.78 %も差がある。また、1 日の中での割合の増減の変化も両日で異なっている。27 日は 20 時台をピークにゆるやかに変化している。それに対し、28 日は、花火大会開始時刻の 19 時台をピークに、特に 19 時前後で急激に変化している。

その地理分布にも違いがある。27 日は全体にまばらに分布しており、時間帯によっての差はほぼ見られない。しかし 28 日は、19 時を中心に、若干ではあるが関東中心部にツイートは集中している。ピークを迎えた 19 時台の地理分布を見ると、花火大会が開催された立川、隅田川、八王子に特にツイートが集まっていることがわかった。19 時台のツイートの地理分布を図 19 に示す。

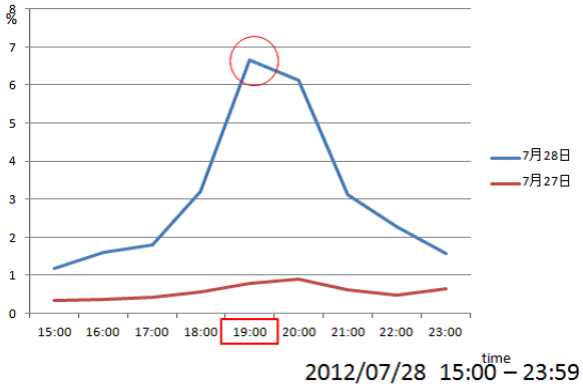


図 17: 「花火」を含むツイートの数の推移

検証結果から、以下のことがいえる。

- イベントの発生時刻と、それに関連するツイートの数には相関関係がある。
- イベント発生地点と、それに関連するツイートの位置には相関関係がある。
- ツイート数、地理分布は短いタイムスパンで変化する。

これらから、ツイート内容・時刻・数・位置情報から、ある地点でのホットトピックの検知が可能であるという仮説を立てた。この仮説に基づいて、システムの設計と構築を行う。

3.4 本章のまとめ

本章では、Twitter と実空間イベントとの相関について、特定のテレビ番組放映時、震災発生時、特定のイベント開催時の 3 つの事例を示した。特定のテレビ放映時の事例では、放映された番組の特定の場面において、Twitter 上で特定のキーワードについてのつぶやきが急増したこと、震災発生時には、時間に伴う人々の行動の推移に伴ってツイート数・ツイート内容が変化していることから、実空間イベントとテキスト内容が、秒単位で相互に影響を受けることが確認された。

特定のイベント開催時については、花火大会に焦点をあて、花火大会の開催日とそうでない日のツイートの動向を比較した。結果、開催日とそうでない日ではツイート数、その地理分布が大きく異なったことから、実空間イベントとテキスト、それに付随する位置情報も相関することがわかった。

以上の事例から、Twitter は時間・位置・ツイート内容において実社会の出来事と相関を持つと考えられる。

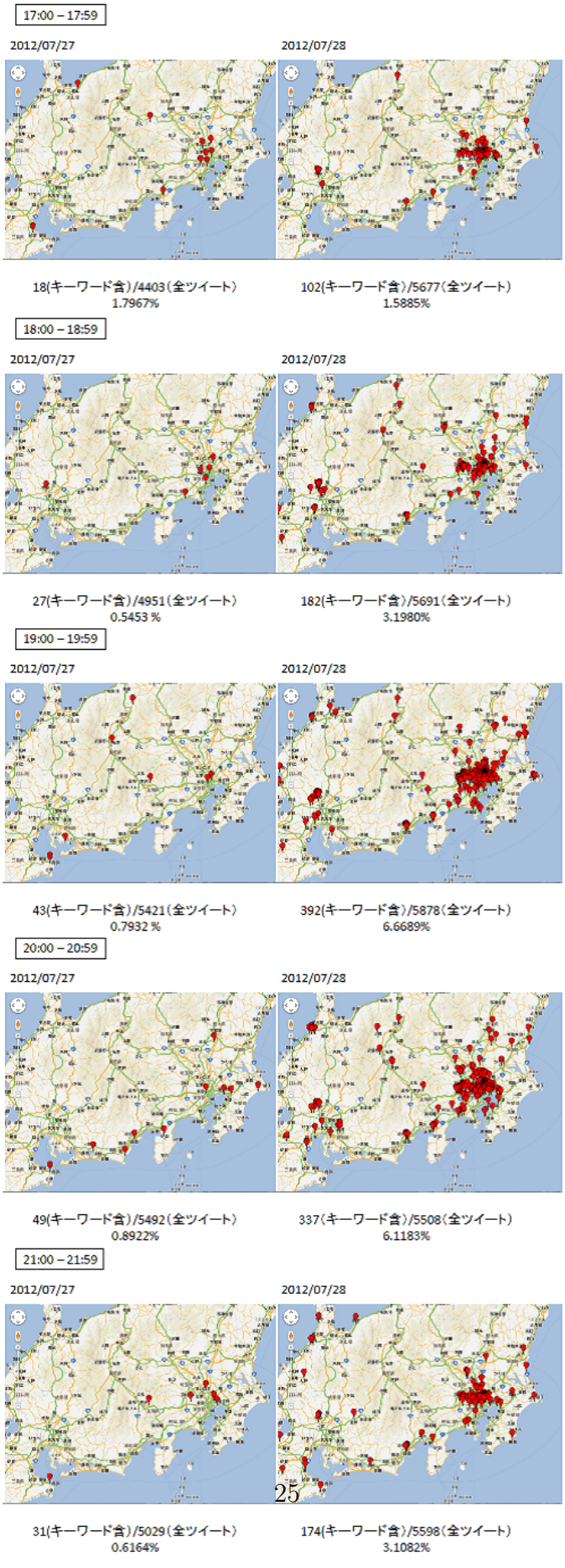


図 18: 「花火」を含むツイート数と地理分布の推移

図 19: 19 時台のツイートの地理分布

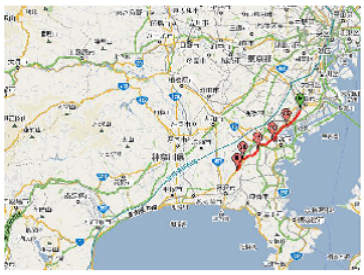
4　関連研究/サービス

本章では、ウェブ上でのイベントやホットトピックの検知を目的とした関連研究、サービスを示す。

4.1　マイクロブログを用いたキーワードと地理的位置の対応付けシステム

梶原 [25] は、ジオタグが付与されたツイートを用いて、言葉と地理的位置との対応付けを行うシステム「最大瞬間なう速システム」を開発した。これは、ジオタグ付ツイートを時間とキーワードで検索を行い、該当するツイートをマップ上に表示するものである。この研究の目的は、キーワードが持つ地理的なあいまい性を除去することで、あるキーワードの地理的なトレンドを明らかにすることである。

例えば、「箱根」というキーワードについて聞き手が想定する場所は、「箱根湯本」、「強羅」、「仙石原」とさまざまである。これを、「箱根」という言葉を含むツイートが、実際にはどのような場所でつぶやかれているかを明らかにすることで、このような言葉のあいまい性を除去することを目指している。また、この研究の中で、「駅伝」のように、時間によって地理的意味が変化するキーワードの存在もわかった。「駅伝」での時間による検索結果の違いと実際の駅伝走者の位置を図 20,21 に示す。



[a]2 区の時間の検索結果　　[b]2 区の時間の駅伝走者の位置

図 20: 検索結果の時間による推移と駅伝走者の位置（2 区）

[a]3 区の時間の検索結果　　[b]3 区の時間の駅伝走者の位置

図 21: 検索結果の時間による推移と駅伝走者の位置（3 区）

4.2 Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors

Sakaki ら [26] は、ジオタグが付与されたツイートの解析を行い、これを利用して地震の検知と発生箇所の推定を行った。

解析は 2 つの段階から成る。

1 段階目として、「地震」「揺れた」といった言葉を含むツイートが、実際の地震の発生直後にされたものかどうかの判定を行う。「地震」や「揺れ」という言葉を含む場合でも、それが本当に地震の発生を意味するものかは定かではない。例えば、一週間前の地震についてのつぶやきである可能性や、「心が揺れる」といった慣用句として使われている可能性も十分にある。これらを正しく判断するため、SVM を用いて有効なツイートかを判定する。

次に、地震直後のツイートの集団からノイズの除去、それらの位置情報から地震発生地点の予測を行う。地震発生時のツイート数の推移を見ると、その数は指数関数的に増加していることがわかる。このことから、ノイズについては時系列的にツイート数を検証し、ノイズか否かを判断する。

位置推定については、カルマンフィルタとパーティクルフィルタ [27] を利用する。

カルマンフィルタとは、直前までの情報と現在の情報を組み合わせることで、現在の状態を推定する手法で、位置推定に広く用いられる。

パーティクルフィルタとは、物体の検出と追跡を同時に行うためのアルゴリズムである。現状態から起こり得る多数の次状態を粒子に見立て、その確立密度から次の状態の予測を行う。

Sakaki らは検知するイベントを地震に絞り、日によって変化する、地震に関する“ホットな場所”を観測することで地震の発生の検知を可能とした。

4.3 Google トレンド

Google トレンドとは、Google.Inc が提供するサービスで、指定したキーワードの被検索数を時系列に表すことで、そのキーワードの人気度の動向を表す。検索数の他に、指定したキーワードと共に検索されたキーワードや、検索された地理位置も見ることができ、さらに検索された地理位置は、時系列ごとに変化する様子を見ることができる。キーワードの盛り上がりを時空間的に観測することができる。

Google トレンドの検索結果例を図 22,23 に示す。

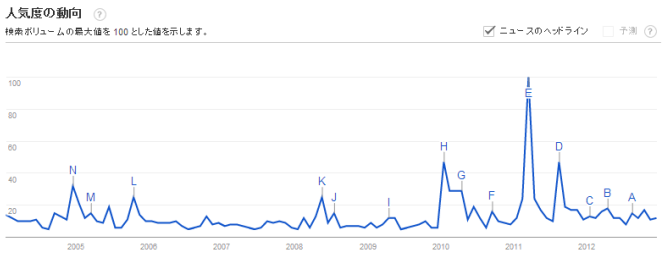


図 22: Google トレンド検索結果例 (1) 検索キーワード : earthquake

図 23: Google トレンド検索結果例 (2) 検索キーワード : earthquake

4.4 関連研究/サービスの比較

上記の 3 つの関連研究およびサービスの総括と比較を行う。どれも時空間解析によって、変化するホットトピックを視覚化および検知を行うという点では共通している。

ここでは、即時性、正確性、ユーザによるイベントの想定の必要性の有無の 3 つの点において比較を行う。

即時性とは、イベントが発生してから検知できるまでの時間の短さを指す。イベント発生から短い時間で検知できる場合に、即時性が高いとみなす。

正確性とは、検索する情報のノイズの少なさを表す。指定したキーワードを含むものの、そのテキスト情報が本当にそのイベントの発生を意味しているとは限らない。キーワードの意味と、そのイベントの発生が合致する情報が多い場合に、正確性が高いとみなす。

ユーザによるイベントの想定とは、あらかじめユーザによるイベントの想定が必要か否か、つまりユーザの趣味・思考に関係なく、社会一般での盛り上がりを抽出できるか否かを指す。

比較結果を表 3 に示す。なお、Sakaki らの研究については、そのシステム名である Toretter と表記する。

表 3: 比較

	即時性	正確性	ユーザによる想定
最大瞬間なう速	○	×	必要
Toretter	○	○	必要
Google トレンド	△	○	必要

即時性

最大瞬間なう速、Toretter は情報源として Twitter を利用している。第 3 章でも示した通り、ツイート数の増減は秒単位で実イベントと対応している。したがって、即時性は高いと言える。反対に、Google トレンドは情報源として被検索数を用いている。イベントが発生してから、人がそれについて検索するまでの時間は人や状況に依存する。

正確性

最大瞬間なう速は、ツイートに形態素解析を行い、その結果とキーワードとのマッチングを行っている。したがって、正確性は形態素解析ツールに依存し、それほど正確性は高くないと考えられる。Google トレンドは、キーワードのマッチングで判断するため、文脈等のあいまい性は発生しない。よって

正確性は高いと言える。Toretter は、解析の第 1 段階で、そのツイートが本当に地震の発生直後にされたものなのかの判断を行っている。正確性は高いと言える。

ユーザによるイベントの想定

最大瞬間なう速、Google トレンドは、ユーザがキーワードを指定する形をとるため、あらかじめイベントの想定が必要である。Toretter は対象を地震に限定しているため、こちらもある程度ユーザが地震の発生を想定している場合に有効となる。したがって、検出されるイベントの種類はユーザ依存となる。

本研究は、時空間解析を用いてテキスト情報からホットトピックの検知を行うという点で、上記の 3 つの研究/システムと共通している。

最も大きな違いは、ユーザによる、イベントの想定が必要か否かという点である。上記の研究/サービスは、キーワード指定等の方法で、ユーザ側であらかじめある程度のイベントの予測が必要となり、ゆえに検知されるホットトピックもユーザ依存となる。しかしこれらの方法では、社会一般で盛り上がっているトピックや、1 番盛り上がっているキーワードを知ることは容易ではない。本研究では、そういったユーザの想定を必要とせず、ウェブ上で一番盛り上がっているトピックをユーザへ提示することを目指す。これにより、ユーザの新たな気付きを促すことができる。

5　アプローチと設計

第 4 章を踏まえ、本章では、本研究でとるアプローチと、その設計について示す。ツイートを緯度、経度、時間、テキストの 4 つの点で解析し、視覚化することで、場所に起因するイベントの検知を行うシステム AKT24 を設計する。まず、システムの設計方針を示し、それを踏まえた概要を述べた後、構成する各モジュールの詳細について述べる。

5.1　アプローチ

第 2 章で示したように、今日、実社会においてインターネット上で個人が発信する声の重要性は高まっている。また、それら一つ一つの情報は、大量に集めることで人の行動にも大きな影響を与えるほどの力を持つことも述べた。これらを踏まえ、インターネット上に存在する大量のテキスト情報および位置情報の管理、解析、視覚化を行うシステムの設計と構築を行う。

解析するテキストおよび位置情報の情報源として満たすべき要件を示す。

- 非常に短いタイムスパンで発信されること
- 大量に収集可能であること
- 各データは匿名性を持つこと

本研究の目的は、“ 今 ” 話題のトピックや場所を検知することである。検知できる話題にリアルタイム性を持たせるために、収集する情報源はより頻繁に更新されることが望ましい。また、2.4 節で述べたように、一つ一つの声は多く集めるほど信頼度は高まる。したがって、できるだけ多くの情報を集めることが必要である。

また、自分の居場所が不特定多数の人々に知られることに対して抵抗を覚える人は少なくない。場合によっては、人の住所や職場などが特定される恐れもある。また、解析するデータの価値が、発信した人の属性等に左右されないためにも、各データは匿名性を持つ必要がある。

以上を踏まえ、利用する情報源としてマイクロブログサービス Twitter を利用する。

5.2　システム概要

本節では、システムの概要を示す。

ジオタグが付与されたツイートを収集し、緯度/経度/時間/テキストの 4 次元で解析を行う。これを地図やグラフで表すことでユーザが視覚的に解析結果を認識できるシステム、AKT 24 の構築を行う。動作の流れを図 24 に示す。

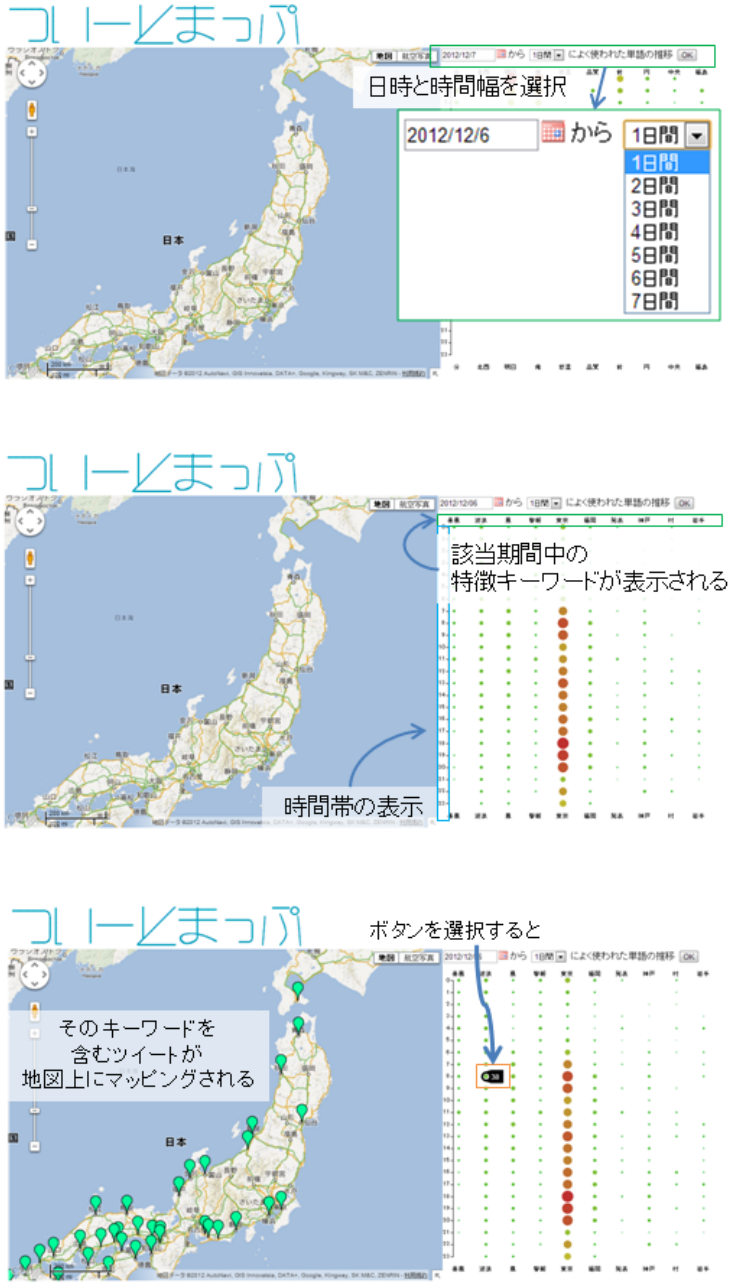


図 24: 動作の流れ

5.3 キーワード抽出手法の検討

設計にあたり、指定日に最も「特徴的」であったキーワードを選出することが必要となる。このキーワード抽出手法の決定にあたり、複数の手法の検討を行った。検討を行った手法は以下の 4 つである。

- 出現回数による選出

- 出現率による選出
- 出現回数の差分による選出
- 出現率の差分による選出

なお、独自にタイムラインデータベースと頻出単語データベースを作成し、ツイートの検索にはこれを用いた。

タイムラインデータベースには、streamingAPI で取得できるツイートのうち、日本をカバーする緯度 127.4414〜148.7109 度、経度 29.9930〜45.8900 度の範囲で投稿されたものを全て格納する。1 日で格納される件数は平均 108261.1 件である。

頻出単語データベースには、タイムラインデータベースへ格納されたツイートの、1 時間ごとの頻出単語（=1 時間の中で出現数の多い単語）が昇順に 1000 個格納されている。

タイムラインデータベースの格納項目を表 4、頻出単語データベースの格納項目を表 5 に示す。キーワードには英数字、記号は含まないものとする。

検討にあたり、各手法ともに 2012 年 10 月 21 日から 27 日までの 7 日間のデータからキーワードの選出を行い、結果を比較した。

表 4: タイムラインデータベース格納項目

user	ユーザ名
date	日時
lat	緯度
lng	経度
text	ツイート内容

表 5: 頻出単語データベース

date	日時
word	キーワード
kensu	1 時間ごとの出現数

5.3.1 手法 1: 出現回数による選出

指定期間中で出現回数の多かったキーワードを抽出する。

指定期間で頻出ワードデータベースへ問合せを行い、結果のキーワードを、その出現件数で昇順にソートする。多いものから 10 選出する。

この手法で抽出されたキーワード 10 つを表 6 に示す。

「笑」、「市」、「明日、「さん」など、多数のキーワードが 7 日間共通して選出された。「さん」、「こと」など、それだけでは意味をなさない単語も抽出された。「タッチ」というワードが多く抽出されているのは、他アプリと連携して利用するユーザが、共通の文面で投稿したためと考えられる。

表 6: 手法 1: 出現回数による選出結果

日	キーワード
2012/10/21	笑、市、明日、県、人、巨人、今日、さん、タッチ、こと
2012/10/22	笑、区、東京、市、県、人、都、こと、明日、今日
2012/10/23	笑、雨、都、駅、店、市、県、明日、さん、こと
2012/10/24	笑、都、明日、店、市、さん、今日、人、こと、県
2012/10/25	笑、宮城、市、県、都、震度、明日、人、こと、今日
2012/10/26	笑、都、明日、市、今日、日、県、人、さん、こと
2012/10/27	笑、市、県、駅、地震、明日、今日、タッチ、人、店

5.3.2 手法 2: 出現率による選出

指定期間中で出現率（＝キーワードを含むツイート数/全体のツイート数）の大きかったキーワードを抽出する。これにより、ツイートの全体数に影響を受けずにキーワードが選出できることが期待される。

まず、指定期間でタイムラインデータベースへ問合せを行い、全体ツイート数を調べる。次に頻出ワードデータベースから、指定期間の頻出ワードを抽出する。それぞれのキーワードごとに、その出現数を全体ツイート数で割り、出現率を算出する。出現率の大きなものを 10 選出する。

表 7: 手法 2: 出現率による選出結果

日	キーワード
2012/10/21	東京、人、明日、今、何、俺、私、前、大阪、分
2012/10/22	東京、人、今、明日、何、俺、分、私、大阪、前
2012/10/23	東京、人、明日、今、波浪、分、何、私、俺、前
2012/10/24	東京、人、明日、今、何、分、私、前、大阪、俺
2012/10/25	東京、人、今、明日、前、俺、何、分、大阪、私
2012/10/26	東京、人、明日、今、大阪、私、何、俺、前、分
2012/10/27	東京、人、明日、今、大阪、何、分、私、前、俺

「東京」、「人」など、共通するワードが多く見られる。9 つの単語が、7 日間共通して選出された。

5.3.3 手法 3: 出現回数の差分による選出

指定期間 T とそれまでとの特異点を見ることで、キーワードの選出を行う。指定日以前の期間のことを指定期間 T' と呼ぶ。指定期間 T' については後に詳説す

る。まず、指定期間内の頻出キーワードデータベースに格納されている 1000 個の各キーワードについて、その出現数を、T/24 時間ごとに算出する。各キーワードについて、指定期間 T' でも同様の処理を行う。これらの値の比較を行い、差分の大きなキーワードを 10 選出する。なお、選出された 10 のキーワードの中で、重複が見られた場合は、差分の大きな方を採用する。イメージを図 25 に示す。

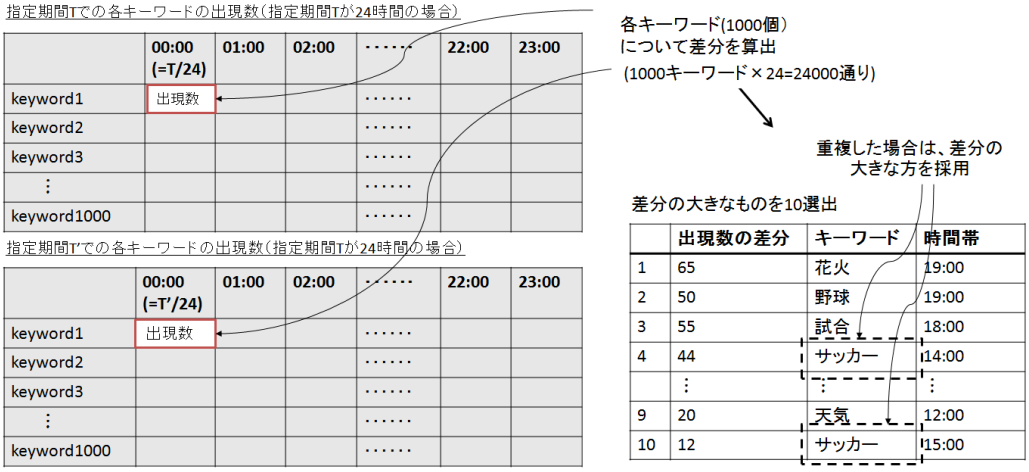


図 25: イメージ図

指定期間 T' の算出手法

ユーザは調べる日と長さを指定する。ここでユーザが指定した日と長さを指定期間 T と呼ぶ。次に、特徴キーワードを選出するにあたり、出現数の比較対象となる期間を設定する必要がある。そこで、ユーザの指定した日から指定した日数さかのぼった日から前日までの期間を指定期間 T' とする。指定期間 T と指定期間 T' のイメージを図 26 に示す。

出現回数の差分による選出結果を表 8 に示す。

「学校」や「仕事」、「事故」など、発信者それぞれの属性、状況が伺える単語が選出された。7 日間で共通して選出された単語は一つもない。また、一般的に多くの人が給料日であると考えられる 25 日には「給料」という言葉が選出されるなど、日に依存するようなワードも見受けられる。

5.3.4 手法 4: 出現率の差分による選出

5.3.3 節で述べた手法を、出現率について同様に調べる。差分の大きなものから 10 のキーワードを選出し、重複が見られた場合は差分の大きい方を採用する。

出現率の差分による選出結果を表 9 に示す。

7 日間共通して選出される言葉は少なく、手法 3 と同様、発信者の状況を表すような単語が多く見受けられた。

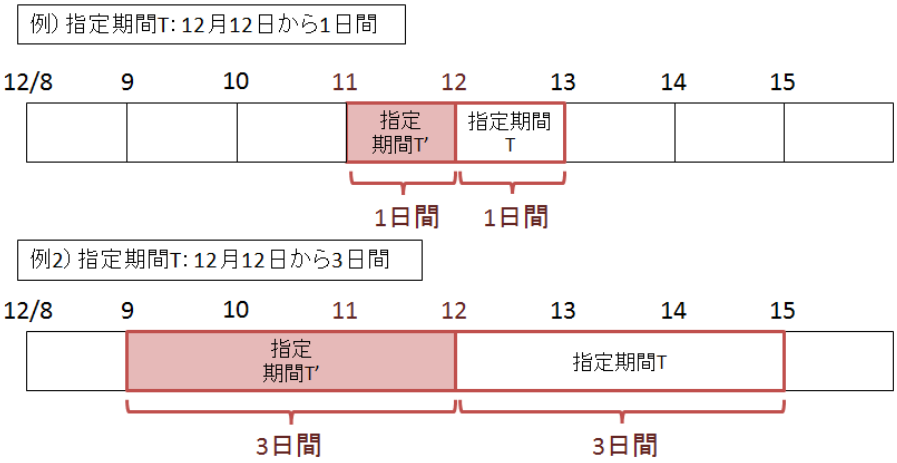


図 26: イメージ図

表 8: 手法 3: 出現回数の差分による選出結果

日	キーワード
2012/10/21	昨日、試験、座、巨人、私、人、方、空、駐車、北九州
2012/10/22	学校、波浪、巨人、授業、病院、仕事、限、開始、沖、中日
2012/10/23	波浪、風、警報、事故、分、品質、付近、静岡、体調、予報
2012/10/24	東京、大阪、明日、沖、中央、円、北海道、名古屋、人、仕事
2012/10/25	阪神、藤浪、大丈夫、氏名、位、給料、沖、分、管野、人
2012/10/26	東京、明日、神奈川、金、祭、宇宙、法則、確認、大阪、携
2012/10/27	東京、人、大阪、今、名古屋、神奈川、投稿、中央、祭、試合

表 9: 出現率の差分による選出結果

日	キーワード
2012/10/21	昨日、座、試験、私、今、人、方、自分、駐車、試合
2012/10/22	学校、仕事、波浪、授業、病院、分、限、沖、円、巨人
2012/10/23	波浪、風、警報、事故、物、市、静岡、予報、分、何
2012/10/24	沖、東京、大阪、円、仕事、中央、発表、福岡、北海道、付近
2012/10/25	限、品質、前、三、出先、俺、構築、和歌山、今年
2012/10/26	東京、明日、宇宙、金、法則、確認、神奈川、祭、私、様
2012/10/27	福島、付近、品質、沖、投稿、試合、名古屋、祭、津波、南西

5.3.5　4 つの手法の比較

上記で示した 4 つの手法の結果を比較し、本研究で利用する手法を決定する。

手法 1、手法 2 で示した方法では、毎日ほぼ同じ単語が抽出された。Twitter は多くの人にとって日常的に利用されるツールで、ユーザは朝の挨拶など、毎日の何気ない一言をつぶやくことが多い。したがって、Twitter 上には日や時間、場所に依存しないで常にツイートされる言葉がある。手法 1、手法 2 で示したような、出現数や出現率を抽出する方法では、こうした常に存在するワードが抽出されてしまう。出現数や出現率が多いというだけでは、その単語が話題となっているかどうかを検知することは難しいことがわかる。

次に、手法 3、手法 4 で示した方法では、7 日間で共通して抽出された言葉はなく、「地震」や「阪神」、「巨人」など、時間や場所に依存すると考えられる言葉が抽出された。また、両手法で検出されるワードはほとんど共通していた。そこで、2 つの手法のうち、より適切なものを選択するため、抽出されたワードの出現数の、時間ごとの推移の比較を行う。なお、比較する日時は 2012 年 10 月 27 日 00 時 00 分から 23 時 59 分までの 24 時間とする。図 27 に両手法の結果を視覚化したものを示す。横軸が抽出されたキーワード、縦軸が時刻、丸の大きさが出現数の大きさを表している。

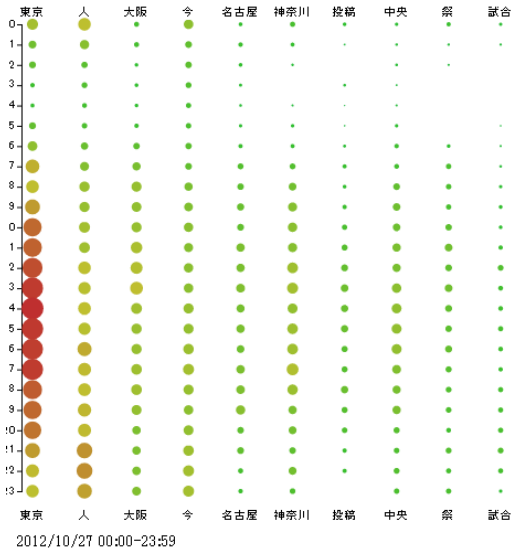
比較すると、手法 3 では数の推移は緩やかで、数の変化に特徴は見受けられなかった。一方で、手法 4 の結果では、複数のキーワードで、瞬間的に出現数が増えている様子を見ることができる。また、瞬間的な増加が見られた単語は、「福島」、「沖」、「津波」であったことから、その時間に福島で地震および津波が発生したと推測することができる。本研究の目的は、「“今”話題の出来事を検知する」ことであるため、キーワードについては、このように短いスパンで変化が認識できるものが抽出されることが好ましい。

以上を踏まえ、キーワードの選出手法は、手法 4 で示した出現率の差分による選出手法を利用することとする。

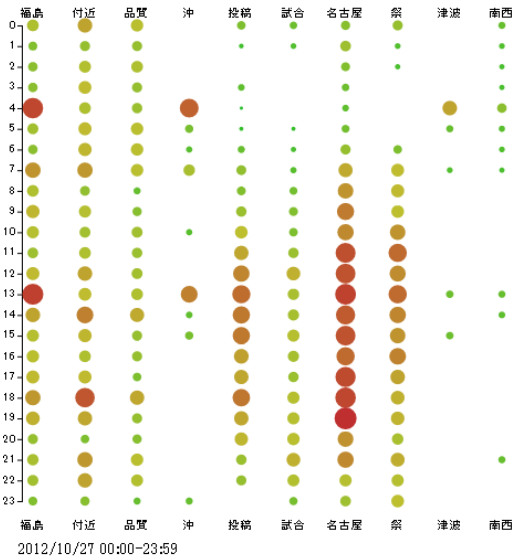
5.4　視覚化手法の検討

前節で、Twitter を利用したイベントの検知には、通常時からの特異点を探すことが有効であると述べた。本システムでは、この特異点を視覚的に捉えることができるユーザインタフェースを目指す。視覚化する情報は、特徴的なキーワードを含むツイートの数とその地理分布とする。

ツイート数の特異点を示すためには、ある程度の長さの時間幅での数の推移を見せる必要がある。時間幅が短すぎると基準となるツイート数が少なくなり、数の推移は時間によるものか、その他の要因によるものか判断できない。反対に時間幅が長すぎると、基準となるツイート数が多すぎるため、特異点を抽出するのが困難になる。



[1] 手法 3



[2] 手法 4

図 27: キーワード選出手法の比較

地理分布については、他に比べてツイートが集中している地域を特異点とする。地理的な特異点は、イベントの性質によってその範囲が異なる。例えば、暴風といった気象情報や、紅葉といった四季のイベントについては、県や地方単位で特異点が抽出される。しかし反対に、大学の学園祭や有名人のコンサートといった、会場が特定して決まっているようなイベントに関しては、非常に狭い範囲が特異点として挙げられる。したがって、地理分布の特異点の抽出にあたっては、イベントによってその範囲に自由度がある必要がある。以上より、視覚化にあたっては以下の方針をとる。

- キーワードの表示は、時間軸に伴うツイート数の推移が見える方式をとる。ユーザが指定可能な項目は、日時と時間幅とする。なお、指定可能な時間幅は 1 日〜7 日間とする。
- 地理的な分布は、キーワードごとに地図上に表示する。特異点の範囲は設定せず、キーワードによってその範囲は変化する。

以上を踏まえ、次節で各モジュールの詳細を示す。

5.5 各モジュール詳細

本システムは以下の 5 つのモジュールで構成する。

- タイムライン取得モジュール

- 頻出キーワード抽出モジュール
- 特徴キーワード抽出モジュール
- ツイート抽出モジュール
- ツイート表示モジュール

タイムライン取得モジュールは常時動作し、Twitter サーバからタイムラインを取得し続ける。頻出キーワード抽出モジュールは一時間ごとに動作し、タイムラインデータベースに格納されているツイートの中での頻出単語の抽出を行う。ユーザからの問い合わせがあると、特徴キーワード抽出モジュールが、頻出キーワードの中から特徴キーワードを選出し、ツイート抽出モジュールにより、特徴キーワードを含むツイートが出力される。特徴キーワード抽出手法については後に詳細する。

システム概要を図 28 に示す。

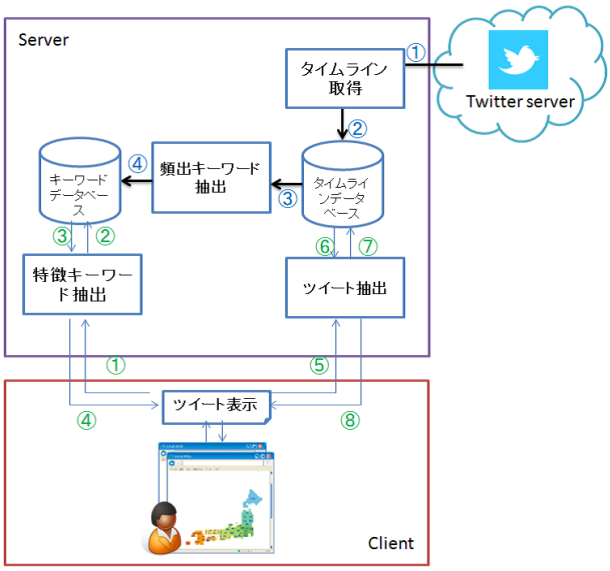


図 28: システム概要図

5.5.1 タイムライン取得モジュール

タイムライン取得モジュールは、Twitter 上のパブリックタイムラインのうち、ジオタグが付与されたものをリアルタイムに取得し、データベースへ格納する。格納する情報は、ユーザ名、ツイートされた日時、緯度、経度、ツイート内容の 5 つである。タイムライン取得モジュールは、常時動作を行う。

5.5.2　頻出キーワード抽出モジュール

頻出キーワード抽出モジュールでは、タイムライン取得モジュールで取得したツイートに形態素解析を行い、1時間ごとに全ツイート中に出現数の多かったキーワードを多いものから1000個抽出し、データベースへ格納する。これらのキーワードを頻出キーワードとする。頻出キーワード抽出モジュールは、1時間に1度動作を行う。

5.5.3　特徴キーワード抽出モジュール

特徴キーワード抽出モジュールでは、ユーザが選択した日時/キーワードでデータベースへ問合せを行い、指定期間内のキーワードデータベースから頻出キーワードを抽出し、それぞれのワードについて、前日までの同時間での出現数との比較を行う。出現数の差分の大きかったものから10個抽出する。特徴キーワード抽出モジュールは、ユーザから問い合わせがあった際、動作する。

5.5.4　ツイート抽出モジュール

指定された期間、さらに時刻内に投稿されたツイート全てをタイムラインデータベースから呼び出し、XML形式にして出力を行う。ツイート抽出モジュールは、ユーザから問い合わせがあった際、動作する。

5.5.5　ツイート表示モジュール

ツイート抽出モジュールで出力されたツイートの中で、特徴キーワードとして抽出されたキーワードを含むツイートのみ選出し、その位置情報をもとにマップ上にピンとして表示する。ユーザから問い合わせがあった際、動作する。

5.6　本章のまとめ

本章では、まず前章までに示した事例や考察から導いた、本研究でとるアプローチと設計方針について述べた。設計方針については、キーワード選出、視覚化の手法を中心に示した。キーワード選出手法については、考えられる4つの手法から適切なものを選択するために検証を行った。その結果、出現率の差分による選出手法が適切であると考えられ、これを利用することとした。

本研究で構築するシステムは、タイムライン取得モジュール、頻出キーワード抽出モジュール、特徴キーワード抽出モジュール、ツイート抽出モジュール、ツイート表示モジュールの5つのモジュールで構成する。

これらを踏まえ、次章では本システムの実装について述べる。

6　実装

本章では、本システムの実装について述べる。section 概要本システムは、twitter streamingAPI、データベース、形態素解析ソフトを利用したツイート解析視覚化ツールである。なお、言語は perl と PHP を使用した。ソフトウェア構成を表 10 に示す

表 10: ソフトウェア構成

OS	MacOS X
web サーバ	Apache2
データベース	MySQL5
形態素解析ソフト	MeCab0.98

6.1　各モジュール詳細

以下に、各モジュールごとの実装環境について述べる。

6.1.1　タイムライン取得モジュール

ツイートの取得には、Twitter 社が提供する streamingAPI を利用する。これにより、世界で発信されるツイートのうち、1/20 を取得することができる。本システムでは日本でのツイートを対象とするため、このうちの緯度 127.4414〜148.7109 度、経度 29.9930〜45.8900 度の範囲内でツイートされたものを取得する。時間による検索を迅速に行うため、データベースへ格納する日時は unix 時間を採用することとし、データベースへ格納する際に通常のタイムスタンプから unix 時間への変換を行う。

データベース格納項目を表 11 に示す。

表 11: タイムラインデータベース格納項目

user	ユーザ名
date	日時
lat	緯度
lng	経度
text	ツイート内容

6.1.2　頻出キーワード抽出モジュール

頻出キーワード抽出モジュールでは、タイムライン取得モジュールで格納されたツイートを 1 時間ごとに取り出し、形態素解析を行う。解析後のツイートを名詞のみ取り出し、出現回数をカウントし、多かったキーワード 1000 個をキーワードと出現数をセットにしてデータベースへ格納する。これを 1 時間に 1 度、毎時間行う。データベースへ格納する項目を表 12 に、頻出キーワード抽出の流れを図 29 に示す。

表 12: 頻出ワードデータベース格納項目

date	ツイートのあった時間帯（1 時間ごと）
word	名詞
kensu	出現数

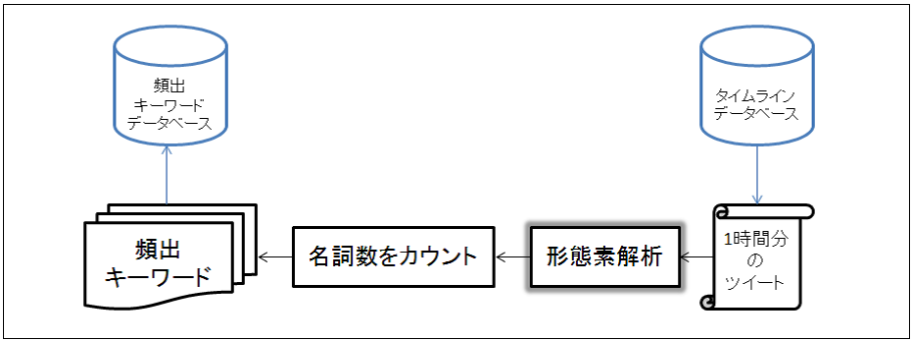


図 29: 頻出キーワード抽出の流れ

6.1.3　特徴キーワード抽出モジュール

特徴キーワード抽出モジュールでは、ユーザが指定した期間で、先ほど作成したデータベースへ検索をかけ、結果として返ってきた各キーワードについて全ツイート中での出現割合を算出する。5.3.3 節で示した指定期間 T' のキーワードについても同様に出現割合を算出し、差分の大きなキーワード 10 つを抽出する。特徴キーワード抽出の流れを図 30 に示す。

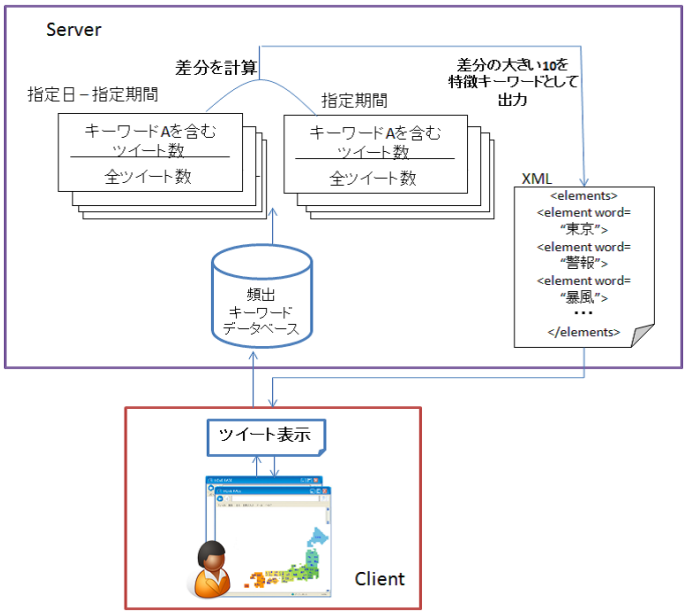


図 30: 特徴キーワード抽出の流れ

キーワード選出アルゴリズム

```
// 指定期間 T と指定期間 T' の設定

for ($i=0; $i < 24*2; $i++) {
    $time[$i]['s'] = ($start-$duration) + $duration/24*$i;
    $time[$i]['e'] = $time[$i]['s'] + $duration/24;
}
// 各キーワード出現数の算出

while ($row = @mysql_fetch_assoc($result)) {
    utf8_encode($row['word']);
    for ($i=0; $i < 24*2; $i++) {
        if ($time[$i]['s'] < = $row['date'] && $row['date'] < $time[$i]['e']) {
        $nWord[$i][$row['word']] += $row['kensu'];
        $total[$i] += $row['kensu'];
        }
    }
}
for ($i=0; $i < 24; $i++) {
    if ($total[$i+24] == 0) continue;
        $wordPrev = $nWord[$i];
        $wordCurr = $nWord[$i+24];

// 指定期間 T と指定期間 T' の出現率の差分の算出

foreach ($wordCurr as $key => $value) {
    if ($total[$i] == 0) continue;
        $word[$key]     +=     ($wordCurr[$key]/$total[$i+24]     -     $word-
Prev[$key]/$total[$i]) * 100;
    }
}
```

6.1.4 ツイート抽出モジュール

ユーザが設定した指定期間 T 内に投稿されたツイート全てをタイムラインデータベースから呼び出す。

URL
http://dora.sfc.wide.ad.jp/topN.php
メソッド
GET
パラメータ
start, end, n

引数の start には、指定期間 T の始まりの時刻を、end には終わりの時刻をそれぞれ UNIX タイムで指定する。n には、表示したいキーワードの数を指定する。本システムでは n=10 とする。

XML 出力例

```
<elements>
  <element word=”風”>
  <entry word=”風” start=”1354741200” end=”1354741350” kensu=”31”/>
  </element>
  <element word=”学校”>
  <entry word=”学 校” start=”1354741200” end=”1354741350”
kensu=”22”/>
  </element>
  <element word=”神奈川”>
  <entry word=”神 奈 川” start=”1354741200” end=”1354741350”
kensu=”26”/>
  </element>
⋮
</elements>
```

次に、画面に表示したキーワードと時間帯が選択された際、呼び出されたツイートの中で、さらに指定した時間帯に投稿されたツイートを選出し、XML で出力する。

URL
http://dora.sfc.wide.ad.jp/getTweet.php
メソッド
GET
パラメータ
start, end

引数の start には、指定した期間の中で、さらに始まりの時刻を、end には終わりの時刻をそれぞれ UNIX タイムで指定する。

XML 出力例

```
<markers>
  <marker date="1354741322" lat="37.9206" lng="139.0598" text="本日の佐渡汽船は悪天候のため欠航。 [pic] http://t.co/v1vLj5mA #eyeland"/>
  <marker date="1354741322" lat="34.2936" lng="134.0633" text="真っ暗〜！ (@ たも屋 林店) [pic]: http://t.co/NOEAr73x"/>
  <marker date="1354741324" lat="36.8436" lng="139.9573" text="寒ー(;O;)"/>
  <marker date="1354741326" lat="35.6702" lng="139.8750" text="あーばかした。本当に寝ちゃったよ すべてノー勉だ"/>
  <marker date="1354741328" lat="37.9223" lng="140.8885" text="仙台の今日の天気 "/>
  <marker date="1354741331" lat="34.6716" lng="135.0304" text="んーやっぱマフラーワンオフしかないかー その前にインチアップかな…"/>
  <marker date="1354741333" lat="38.0814" lng="138.4374" text="朝食中〜 風が強えー !! "/>
  <marker date="1354741337" lat="35.5371" lng="139.6351" text="@quebon_1377 おはようございます (^O^) ／"/>
⋮
</markers>
```

6.1.5 ツイート表示モジュール

ツイート抽出モジュールで出力されたツイート全てについて、キーワードを含むか否かの判定を行う。キーワードを含むツイートのみを抽出し、その位置情報から地図上にピンとして表示する。

アルゴリズム

```
// 指定時刻にあった全ツイートについて参照する
for (var i=0; i<tweets.length; i++){
    var date = tweets[i].getAttribute(”date”);
    var text = tweets[i].getAttribute(”text”);
    var point = new google.maps.LatLng(
    parseFloat(tweets[i].getAttribute(”lat”)),
    parseFloat(tweets[i].getAttribute(”lng”)));
// キーワードが含まれているものを選出、マップ上に表示
    if (text.indexOf(word) >= 0) {
        var icon = ’http://chart.apis.google.com/chart?chst=d_map_pin_letter
_chld=|00fa9a|000000’;
        var marker = new google.maps.Marker({
            map: map,
            position: point,
            icon: icon
        });
        bindInfoWindow(marker, map, infoWindow, text);
        markers[n][j++] = marker;
    }
}
```

6.2 本章のまとめ

本章では、設計に基づく実装環境について、各モジュールごとに示した。また、キーワード選出アルゴリズムについても述べた。

7　評価

本システムの目的は、世の中のホットトピックを、Twitter へ投稿されたテキストと位置情報を用いていち早く検知することである。これらについて評価を行う。評価項目は実用性の有無とイベント検知の正確さの 2 点とし、検知にかかる時間と本システムによって検出されたキーワードの正確性について検証を行う。

7.1　実用性の有無

本システムのイベント検知にかかる時間を測定し、この速度をもって実用性の評価を行う。

- キーワードデータベースの作成にかかる時間
- 検索にかかる時間

の 2 つの時間を計測し実用に耐えうるものかについて評価を行った。

キーワードデータベースは、毎時間 1 回作成される。作成にかかる時間については、1 時間ごと 24 時間分の平均を算出した。結果を表 13 に示す。

検索にかかる時間については、7 日間分の平均を算出した。結果を表 14 に示す。

結果、データベース作成にかかる時間は 24 時間平均で 10.375 秒、検索にかかる時間は 7 日間平均で 8.72 秒と、どちらも 10 秒前後であることがわかる。これより、本システムは実用に耐え得るものであると考えられる。

表 13: キーワードデータベース作成にかかる時間

時間	所要時間
2012/12/18 00:00	16sec
2012/12/18 01:00	12sec
2012/12/18 02:00	8sec
2012/12/18 03:00	8sec
2012/12/18 04:00	7sec
2012/12/18 05:00	6sec
2012/12/18 06:00	7sec
2012/12/18 07:00	8sec
2012/12/18 08:00	10sec
2012/12/18 09:00	10sec
2012/12/18 10:00	10sec
2012/12/18 11:00	10sec
2012/12/18 12:00	6sec
2012/12/1813:00	12sec
2012/12/18 14:00	11sec
2012/12/18 15:00	5sec
2012/12/18 16:00	11sec
2012/12/18 17:00	11sec
2012/12/18 18:00	12sec
2012/12/18 19:00	13sec
2012/12/18 20:00	11sec
2012/12/18 21:00	14sec
2012/12/18 22:00	15sec
2012/12/18 23:00	16sec
平均	10.375sec

表 14: 検索にかかる時間

日	所要時間
2012/10/21	8.77sec
2012/10/22	8.85sec
2012/10/23	8.23sec
2012/10/24	9.06sec
2012/10/25	8.49sec
2012/10/26	8.90sec
2012/10/27	8.73sec
7 日間平均	8.72sec

7.2　イベント検知

本節では、本システムのイベント検知の性能についての評価を行う。

本システムによりイベントの発生とその発生箇所が視覚的に明らかとなったケースのうち、特に特徴的であったものを示す。

7.2.1　2012 年 10 月 13 日

この日表示された 10 のキーワードは、「試合」、「香川」、「戦」、「花火」、「空」、「川島」、「綺麗」、「勝利」、「福井」、「代表」であった。この日のキーワード出力結果を図 31 に示す。時刻に伴うツイート数の推移を見ると、18 時台に「花火」、16 時台に「空」を含むツイートが急増していることがわかる。

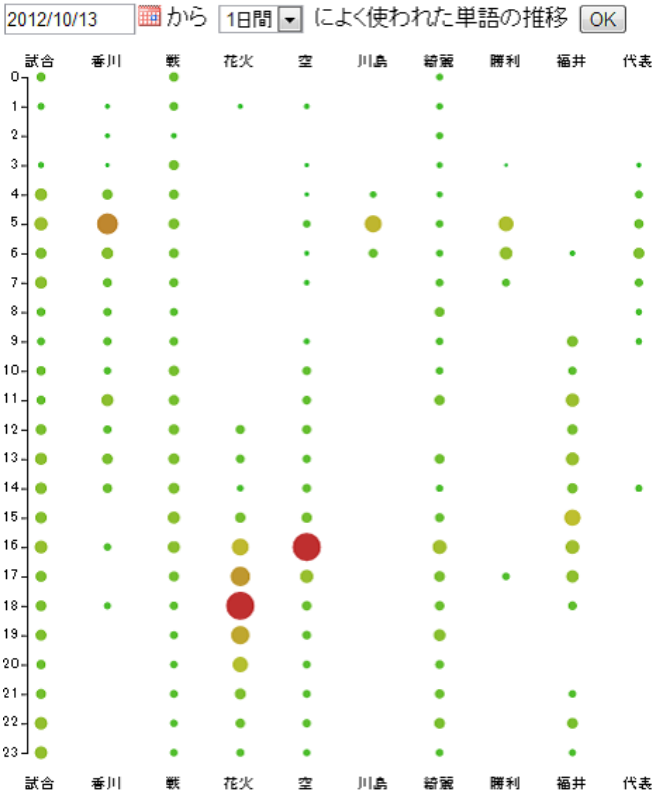


図 31: 2012 年 10 月 13 日キーワード出力結果

キーワード「花火」

キーワード「花火」を含み、18 時台に投稿されたツイートの地理分布を表示させた様子を図 32 に示す。日本全体で見て、関東にツイートが集中していることが

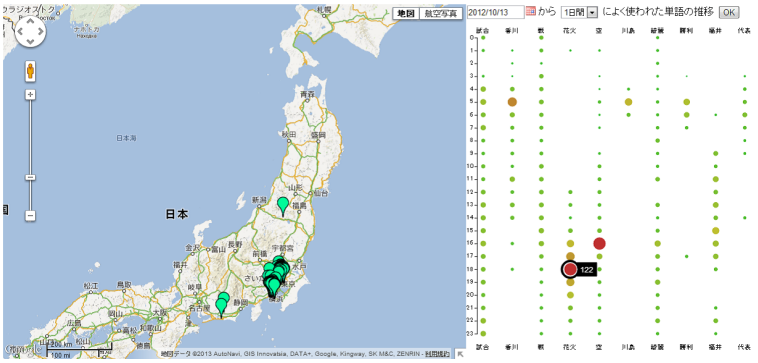
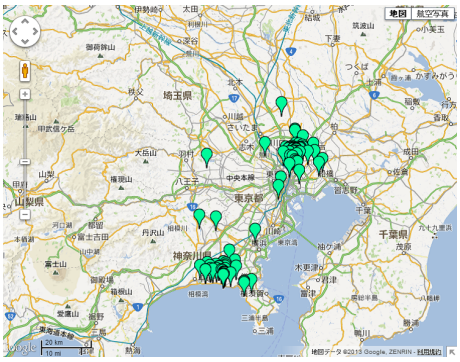


図 32: 2012 年 10 月 13 日、キーワード「花火」の出力結果

わかる。図 32 の地図の縮尺を大きくすると、大きくわけて 2 つの地点にツイートが集中している様子を見ることができる。さらにこの日の花火大会の開催状況を調べると、東京都足立区 [28]、神奈川県藤沢市江の島 [29] でそれぞれ花火大会が開催されていたことがわかった。どちらの花火大会も開始時刻は 18 時である。図 32 で示した地図の縮尺を大きくしたものを図 33[a] に、同日に開催された花火大会の打ち上げ箇所を示した地図を図 33[b] に示す。



[a] 出力結果拡大

[b] 花火大会の打ち上げ箇所

図 33: 2012 年 10 月 13 日の出力結果と同日の花火大会打ち上げ箇所

花火大会の開始時刻である 18 時にツイート数が増えていること、ツイートの地理分布が明らかに打ち上げ地点に重なっていることから、本システムによって花火大会の開催時間、開催地点を検知することができたと考えられる。

キーワード「空」

次にキーワード「空」を含み、最も数の多かった 16 時台に投稿されたツイートの地理分布を図 34 に示す。ツイートが関東に集中している様子がわかる。ここで、

ツイートの内容を抜粋すると「空がすごい」、「空の雲の形が不気味である」、「西の空がいわゆる一つの地震雲 http://t.co/MPdhbaT5」、「こんにちは！空がいい感じです〜(ˆ_ー) ー☆ http://t.co/pT2pFeF6」などと、空に浮かぶ雲の様子についてのツイートが多く、空の写真を添付しているものもみられた。投稿された画像から、空の現象について皆共通のツイートしていると考えられる。ここから、調査日に関東地方で空になんらかの異変が発生したと予測が立てられた。調査を行うと、この日見られたのは波状雲という特徴的な雲であったことがわかった。各種サイト [30] にも、この日の空についてまとめられている。図 35 にツイートに含まれていた画像例を示す。

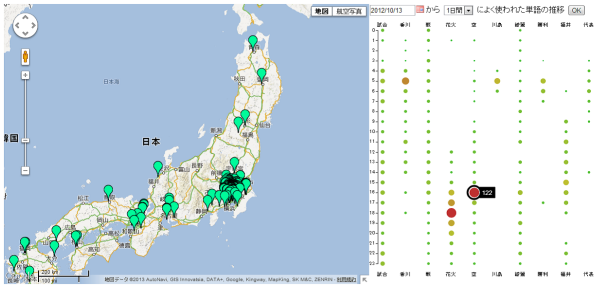


図 34: 16 時台に投稿された、「空」を含むツイートの地理分布

図 35: ツイートに含まれていた画像例

7.2.2　2012 年 11 月 18 日

表示された 10 のキーワードは、「人」、「風」、「綺麗」、「中央」、「昨日」、「投稿」、「紅葉」、「動物」、「法」、「拡大」であった。そのうち、「紅葉」を含むもので、特に

ツイート数が多かった 12 時台から 15 時台の地理分布を図 36 に示す。ツイートは、関東から九州地方に分布し、北海道、東北地方には 1 つもピンがたっていない。

図 37 は、民間の気象予報会社 [31] が発表した 2012 年の紅葉見頃予想マップである。2012 年 11 月 18 日から 1 週間以内に見頃を迎える地点に黒い印をつけた。この図から、検証を行った 2012 年 11 月 18 日には、北海道/東北地方では既に紅葉が終わっていたと考えられる。

以上から、調査日時点での紅葉の色づき状況が検知され、またその分布も現実に即したものであったと評価できる。

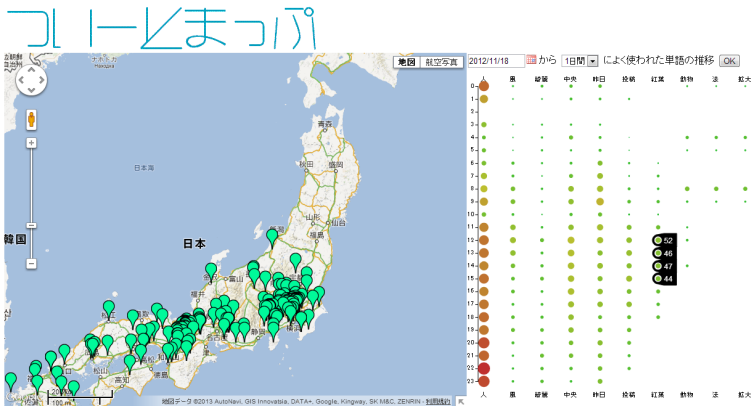
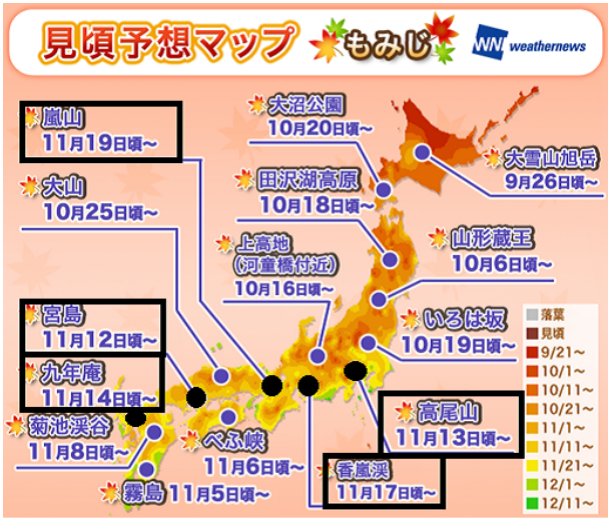


図 36: 2012 年 11 月 18 日



出所)weather news, 紅葉見頃予想,http://weathernews.com/ja/nc/press/2012/120912.html

図 37: 2012 年紅葉見頃マップ

7.2.3　2013 年 1 月 2 日

2013 年 1 月 2 日に検知されたキーワードは「明日」、「二」、「箱根」、「俺」、「度」、「艸」、「恒」、「波浪」、「酒」、「付近」であった。このうち、「箱根」を含むツイートの地理分布とその時間に伴う推移を図 38 に示す。分布が線状に広がり、さらに時間に伴ってその位置も変化している。箱根駅伝の 1 区から 4 区までのコースと比較すると、これとほぼ同時に変化している様子を見てとることができる。時間に伴うイベント位置の変化を検知したと考えられる。

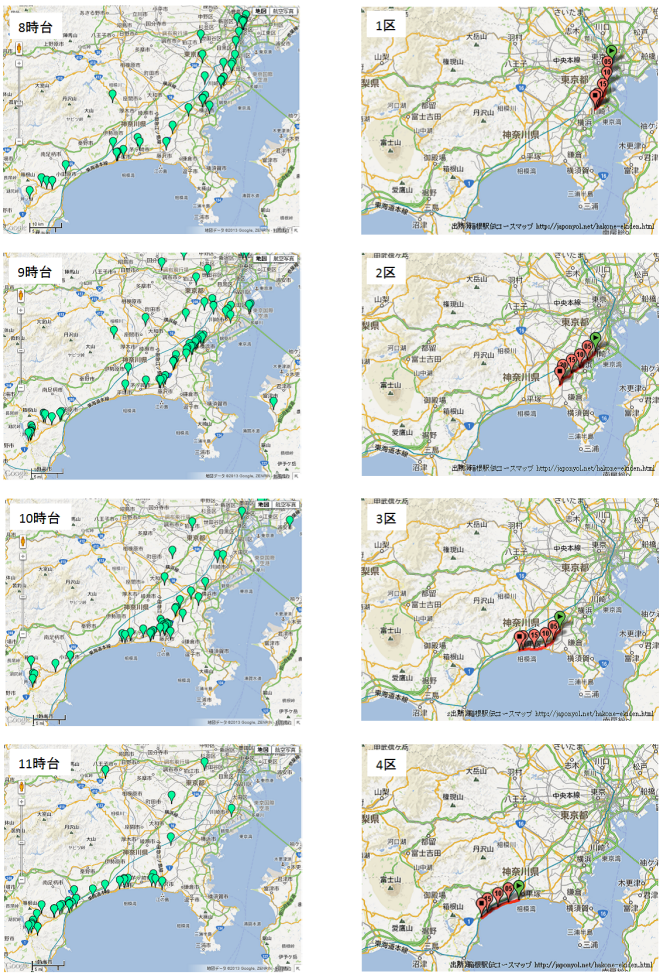


図 38:「箱根」を含むツイートの地理分布の時間別推移と箱根駅伝 1 区から 4 区

7.2.4　2012 年 11 月 10 日

この日検知されたキーワード「東京」、「展示」、「波浪」、「投稿」、「祭」、「東」、「大阪」、「試合」、「南」、「静岡」のうち、「展示」を含むツイートの地理分布を図

39 に示す。りんかい線国際展示場駅付近に多く分布している。1 日の中で 9:00 から 17:00 までのツイート数が多いことから、この時間帯になんらかのイベントが発生していたと考えられる。ツイートの内容は、「東京国際展示場なう」「I'm at 東京国際展示場」など、そのイベント内容についてのツイートが少なく、ここからイベント内容について知ることはできなかった。

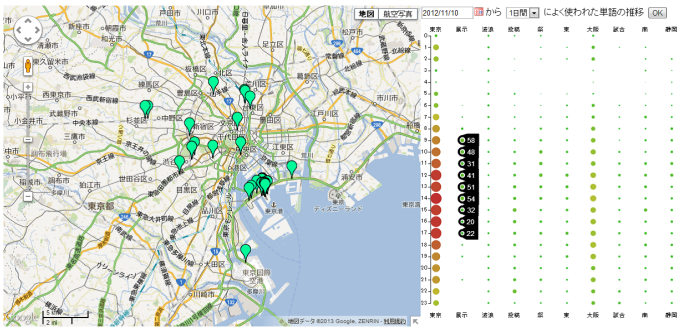


図 39: 「展示」を含むツイートの 9:00 から 17:00 までの地理分布

7.3 利用者からの声

2012 年 11 月 22、23 日に東京ミッドタウンで開催された OpenResearchForum[32] に本システムを出展し、来場者に実際に利用してもらった。この時利用者から多く出た意見を以下に記す。

- 検知できるイベントのジャンルを絞る/分ける

地震、天気の異変など、検知する事柄のジャンルを絞ることで、現在のものよりも精度の向上や実用性の向上を図ることができるのでは、という意見が出た。しかし本システムの特徴の一つとして、「検出されるイベントの種類はユーザの想定を必要としない」という点がある。検知できるイベントの対象を絞らないことでユーザの新たな発見を促すことができ、この点に本研究の有意性があると考える。

- ツイート同士の相関

現行のシステムでは言葉と位置を紐づけるのみで、ツイート同士の相関については検討されていないという点が指摘された。ツイート同士を、その位置や内容でクラスタリングすることでさらに精度を高めることができると考えられる。ツイートのグルーピングや、その意味的距離を測る研究は多く行われており、本研究に関しても今後検討されるべき項目である。

- イベントの予測

現行のシステムは、今実際に発生している、または発生したイベントを検知する。しかし、ユーザが本当に知りたいのは「これから何が起こるのか」である、という声があった。ツイートを用いて未来の予測を行う方法として、機械学習の手法を取り入れることが考えられる。しかし、Twiter に投稿されるテキスト情報は記号やスラングなど特定の人にしかわからない言葉も多く、この中からパターンを見つけ出すことは非常に困難で実現性は低いと考えられる。

- プライバシーに関して

大部分の人が、自分のツイートがフォロワー以外に読まれることに対して抵抗感を抱くことがわかった。また、streamingAPI を用いるという安価な方法で、誰でも大量のツイートを収集することができるという事実を知らなかったユーザも多い。ツイート内容だけでなく位置情報も取得できることから、「今後ジオタグの使い方を改めたい」という意見も聞かれた。

本システムを実装するためには大量の非構造データが必要であり、ユーザがテキストや位置情報を発信する際に感じる抵抗感をできるだけ減らす必要がある。したがって、プライバシーへの配慮は不可欠であり、この点についてさらなる検討が必要である。

- その他

その他の意見として、「ツイート発信者のランク付けを行う仕組み」の提案があった。アカウントごとに過去のツイートの分析を行い、その有用性の判定を行うことで、実現は可能であると考えられる。しかしそれには、ツイート収集、プライバシーへの配慮の 2 つの点で困難が伴う。

ツイート収集に関して、本システムで利用している streamingAPI で取得できるツイートは、投稿された全ツイートのうち 1/20 である。そのため、1 つのアカウントのツイート全てを取得することができず、偏りなく有用性の判定を行うことが難しい。また、アカウント数は膨大であるため、アカウントごとにツイートを分別するには時間がかかる。

また、先に記述したように、多くのユーザが自分のツイート/位置情報が第三者に読まれることに対して抵抗を感じており、ひとつひとつの情報に匿名性を持たせることが不可欠である。アカウントごとにツイートを管理すると、その匿名性が弱まり、プライバシーへの配慮に欠けた設計となってしまう。

7.4 考察

7.2 節で行ったイベント検知についての評価に基づき、考察を行う。

検知できたキーワードについて、本システムの出力結果から、各イベントが発生したと考えられる時間幅、イベント内容、地理範囲について表 15 にまとめた。表を見ると、イベントの発生が考えられる時間幅や地理範囲はそれぞれ異なることがわかる。そこで、時間幅や地理範囲から zone 1 から zone4 まで設定し、それぞれのキーワードについて振り分けた。zone1 から 4 までの特徴は以下の通りとする。振り分けた図を図 40 に示す。

- zone1:発生したと考えられる時間幅は長く（3-5 時間）、地理範囲は狭い。
- zone2:発生したと考えられる時間幅は短く（1 時間単位）、地理範囲は狭い。
- zone3:発生したと考えられる時間幅は短く（1 時間単位)、地理範囲は広い。
- zone4:発生したと考えられる時間幅は長く（3-5 時間）、地理範囲は広い。

表 15: 比較

検出した キーワード	ツイート数が 盛り上がった時間	発生したと 考えられるイベント内容	地理範囲
花火	18:00	花火大会	花火打ち上げ箇所周辺
空	16:00	雲の異変	関東
紅葉	12:00-15:00	紅葉状況	関東から西日本
箱根	8:00, 9:00, 10:00, 11:00	箱根駅伝	第 1 区から 4 区までの 各走者の位置とその付近
展示	9:00-17:00	×	国際展示場付近

zone1 に振り分けられた「展示」は東京国際展示場を指す場合が多く、イベント事が開催されると考えられる土日祝日に頻繁に検知された。このゾーンに振り分けられるキーワードは、瞬間的なツイート数の増減はあまりなく、3 時間から 5 時間程度、恒常的に盛り上がりが検知されたことから、検知されるイベントは突発的に発生するものではなく、長期間に渡って発生する事象であることがわかる。また、地理分布は狭い範囲に集中していることから、地名やランドマークとなる建物を指すことが多い。以上から、zone1 に振り分けられるイベントは、イベント会場などの特定の場所であらかじめ予定されていたコンサートや握手会といったイベントが多いと考察される。

zone2 に振り分けられるのは、突発的に、局所的に発生する事象である。時間によって検知される場所が変化するキーワードも zone2 とする。このゾーンには、「箱根（箱根駅伝)」が振り分けられる。検知できた地点の範囲が狭く、そのため

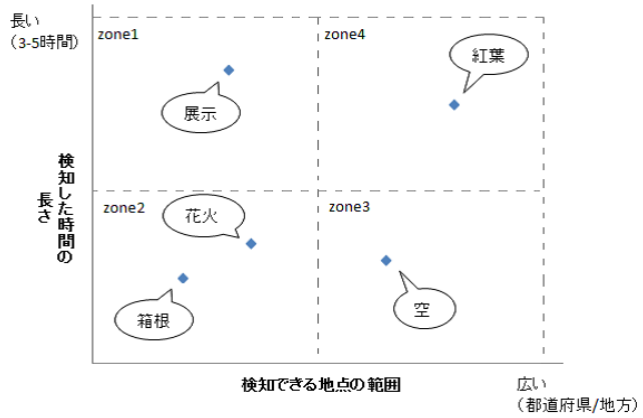


図 40: 判定されたイベントの分類

zone1 の結果と同様、検知できたのは人為的なイベント事が多かった。その他、「渋滞」がこのゾーンにあてはまる。

zone3 に振り分けられるイベントは、突発的に、広い範囲で発生するという性質を持つ。そのため、人為的なイベントが検知されにくく、空の異変や地震といった自然現象によるものが多かった。

zone4 に振り分けられるイベントは、発生時間が長く、さらに広範囲で発生するという性質を持つ。zone3 と同様、人為的なイベントは検知されにくく、このゾーンに振り分けられるイベントもまた、自然発生的なものとなった。zone3 との違いは、その発生時間の長さである。地震や空の様子は、一日の中で移り変わるのに対し、紅葉などの自然の中で起こる現象は、数日/数週間といった長いスパンで変化していく。

以上の結果から、本システムによって、コンサートや花火大会などの人為的なイベントから、気象情報など幅広い分野に渡ってイベントの検知ができることがわかった。また、イベントの性質によって検出のされ方が異なるということもわかった。

以上を踏まえ、次章では本研究で行った検証や構築したシステムについてまとめ、今後について述べる。

8　結論

本章では、本研究のまとめと今後の課題について述べる。

8.1　本研究のまとめ

本研究ではマイクロブログサービスである Twitter と実空間とのつながりに注目し、Twitter に投稿されるテキスト情報と位置情報を利用することで、“ 今 ” 話題となっている物事とそれに付随する場所を検知する手法を提案した。

手法の提案にあたり、Twitter と実空間で発生したイベントとの間の相関について調べるため、1. 東日本大震災発生時、2. 花火大会開催日の 2 つの例について、その日投稿されたツイートの動向を独自に検証した。検証の結果、東日本大震災発生時には、震災発生からの時間の推移にともなって、ツイート数、ツイート内容が変化していくことがわかった。また、花火大会開催日には、花火に関するツイートが通常時に比べて大幅に増加し、その地理分布も花火大会会場に集中していることがわかった。以上の結果から、Twitter 上の情報と実空間イベントは相関すると判断し、Twitter 上のツイートから実空間で発生したイベントの検知を行うシステム、AKT24(Ayako Kurata Tweet-analyzer 24h) の実装を行った。AKT24 はツイートの収集/解析を行い、“ 今 ” 話題のキーワードの選出と、時間によるその出現率の推移、地理分布を視覚化することで、話題となっている事象とその場所をユーザに提示する。キーワードの選出については複数の手法を比較・検討し、通常時との出現率の差分を利用する手法を採用することとした。

評価の結果、AKT24 により、人為的なイベントから気象状況の異変までさまざまな実空間上の事象について検知することができた。また、検知できるイベントは発生した時間の長さと地理範囲の広さから大きく 4 つに分類され、それぞれ検出のされ方が異なることがわかった。

8.2　今後の課題

8.2.1　イベント発生箇所の定量的検知

現行のシステムでは、ツイートの地理分布を地図上への表示することで、ツイートの過密地域およびイベント発生箇所の判定をそれぞれのユーザに委ねている。これでは、正確に場所を判定することが難しく、各イベントの地理依存度も判定しづらい。

ツイート同士の距離を測定し、最も過密度の高い範囲を数値的に示すことができれば、イベント発生地点を正確に知ることができ、またその範囲の広さから各イベントの地理依存度も知ることができる。

8.2.2　プライバシーへの対処

ツイートを利用するにあたり、各ツイート発信者のプライバシー保護への対応にはまだ検討の余地があると考えられる。現在は、各ツイートについてその位置とツイート内容を全て紐づけて閲覧できるようになっているが、これにより個人やその居場所が特定される可能性も否定できない。距離の近いツイートをグループ化し、グループごとにまとめてツイート内容を提示するなど、個人の特定を防ぐ方法はいくつも考えられる。

SNS などの普及によって個人の発信力が高まっていくのに伴い、それを収集し、他の目的に役立てる試みは今後も増えていくと考えられる。このような試みにおいて、個人が発信したデータをどのように扱い、守るかといった問題は重要であり、本システムに限らず検討されるべき事項である。

参考文献

[1] 株式会社東急レクリエーション　ranKin ranQueen http://www.ranking-ranqueen.net/index.html

[2] Facebook, http://facebook.com/

[3] mixi, http://mixi.jp/

[4] LinkedIn, http://jp.linkedin.com/

[5] twitter Inc, twitter http://twitter.com

[6] 財団法人インターネット協会,「インターネット白書 2012」(2012)

[7] ReTweeter!, http://retweeter.unicco.in/

[8] バイラッター, https://jp.twitter.com/viratter

[9] hashtagsjp, http://hashtagsjp.appspot.com/

[10] SEO Japan,「【2012 年版】ツイッターの歴史と現状が 1 枚の絵でわかるインフォグラフィック」,2012 年 3 月 1 日, http://www.seojapan.com/blog/twitter-infographic-2012

[11] 酒巻智宏, 岩井将行, 瀬崎 薫:マイクロブログのジオタグを用いたユーザの行動パターンの調査に関する研究, 全国大会講演論文集,2011(1),pp.787-789,(2011)

[12] foursquare, foursquare https://ja.foursquare.com/

[13] 株式会社ライブドア, ロケタッチ　http://tou.ch/

[14] Google Inc, Google Latitude http://www.google.com/intl/ja/mobile/latitude/

[15] Casey Barto (2011) ,Tips for Location-Based Marketing, Tips for Location-Based Marketing

[16] 日立製作所,「ユーザーの欲しい情報を希望する時間に携帯電話の待ち受け画面へ　画像付き情報として配信する携帯情報サービス「キメクル」を本格的に開始」　http://www.hitachi.co.jp/New/cnews/month/2005/08/0808.pdf

[17] 飯尾 淳, 吉田圭吾, 小池亜弥,　清水浩行, 白井康之,　桑山晃一, 栗山桂一, 小浪宏信, 高山隼佑:属性付き位置情報ログが示す行動特性と消費傾向の関係, 情報処理学会論文誌,Vol.52 No.7 pp.2256-2267,(2011)

[18] 株式会社コロプラ　http://pc.colopl.jp/pages/wl/welcome.html

[19] 株式会社コロプラ　プレスリリース「コロプラ、位置ゲープラットフォーム「コロプラ」のユーザ数が 300 万人を突破！」2012 年 10 月 3 日 http://colopl.co.jp/news/pressrelease/2012100301.php

[20] ITMedia mobile「ケータイゲーム初!「コロプラ」の物産展が大盛況」2011 年 6 月 10 日　http://www.itmedia.co.jp/mobile/articles/1106/10/news025.html

[21] 日経 BP 社　ITpro　「ネット上で急増する「非構造化データ」ビッグデータの活用がビジネスを制す」,2011 年 9 月 30 日, http://itpro.nikkeibp.co.jp/article/COLUMN/20110921/369109/

[22] 日本テレビ, 金曜ロードショー「天空の城ラピュタ」, http://www.ntv.co.jp/kinro/lineup/20111209/index.html

[23] 「「ラピュタ」に毎秒 2.5 万ツイート――TV、SNS と間合い探る（ネット新潮流）」『日本経済新聞』2011 年 12 月 20 日

[24] 日本テレビ, 新世紀エヴァンゲリオン,http://www.ntv.co.jp/pre/evangelion/

[25] 梶原 浩紀:マイクロブログを用いたキーワードと地理的位置の対応付けシステム, 卒業論文 (2010)

[26] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo:Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,19th International World Wide Web Conference,pp.851-860(2010)

[27] 加藤 丈和:パーティクルフィルタとその実装法, 情報処理学会研究報告 [コンピュータビジョンとイメージメディア]2007(1),pp.161-168(2007)

[28] 足立区観光交流協会, あだち観光ネット,http://adachikanko.net/hanabi/index.html

[29] 公益社団法人 藤沢市観光協会, 藤沢市・湘南江ノ島,http://www.fujisawa-kanko.jp/event/fujisawahanabi.html

[30] NHN Japan Corp., 「10 月 13 日に話題になったシマシマの雲は波状雲」,http://matome.naver.jp/odai/2135012341675833701

[31] WEATHER NEWS,INC, http://weathernews.com/

[32] 慶應義塾大学,OPEN RESEARCH FORUM 2013,http://orf.sfc.keio.ac.jp/

謝辞

本論文執筆にあたり、ご指導賜りました慶應義塾大学環境情報学部教授 村井純博士、同学部教授 中村 修博士、同学部准教授 楠本博之博士、同学部専任講師 Rodney D. Van Meter Ⅲ博士、同学部准教授 三次 仁博士、同学部教授 武田圭史博士、同大学 DMC 機構専任講師 斉藤賢爾博士に感謝致します。

また、常にご指導頂き、たくさんのことを学ばせてくださった慶應義塾大学環境情報学部准教授 植原啓介博士、シンガポール国立大学 佐藤雅明博士に感謝致します。右も左もわからぬまま研究室に飛び込んだ私に様々な助言や機会を与え、見守って下さいました。お 2 人のご指導がなければ、このチャレンジに満ちた 3 年間を送ることはできませんでした。この研究生活は私にとって、とても濃い、意味のある期間でした。ありがとうございました。

慶應義塾大学村井研究室 木本瑞希氏、三條場直希氏、同大学政策・メディア研究科 鈴木詩織氏、OB/OG である波多野敏明氏、Do Thi Thuy Van 氏、東京大学大学院 鵜飼 祐氏、MIT メディアラボ 澤田 暖氏、京都大学大学院 村上滋希氏に感謝致します。皆さまからは勉強だけにとどまらず様々な場面でアドバイスを頂き、たくさんのことを学ばせて頂きました。皆さまからの助言や励ましを胸に、今後も邁進していきます。

また、長い間成長を共にし、大学ではそれぞれの道に歩みながらも折々で助けてくれた成瀬台中学校 OB/OG の皆さん、八王子東高等学校 OB/OG の皆さんに感謝致します。皆さんと過ごした日々は、大学生活同様、今の自分を作る大切なものです。

最後に、生まれてから現在に至るまで、私の意思を認め、常に支えてくれた父 和彦氏、母 節子氏、姉 陽子氏に感謝致します。諸氏の存在がなければ、上記の方々と素晴らしい時間や経験をともにすることもできなかったでしょう。ありがとうございました。

以上をもって謝辞とさせていただきます。