E-04

地域イベント発見のためのジオタグ付マイクロブログを用いたノーマルパターン検出手法

Construction of Usual Behavior Patterns of Crowds for Twitter-based Geo-social Event Detection

藤坂 達也† Tatsuya Fujisaka　李 龍‡ Ryong Lee　角谷 和俊‡ Kazutoshi Sumiya

1．はじめに

Twitter [5]に代表されるマイクロブログの急速な普及により，人々は容易に思いや感情，さらに体験までも即座に共有することが可能となった．また，ディスクトップ PC だけでなく iPhone のようなスマートフォンからの発信も増加している．Twitter の場合，通常はコメントに位置情報と発信時間が付与されるが，位置情報はあらかじめ手動で設定されることが多い．しかし，スマートフォンからの発信において，GPS 機能を用いて自動で現在地の緯度，経度がマイクロブログに付けられることも可能となり，誰が，どこで，いつ，どんな情報を発信したか分析することができる．

我々は，このようなマイクロブログを**ジオタグ付きマイクロブログ**と定義し，このジオタグ付マイクロブログを分析することで，様々な場所で起きている日常とは異なる地域社会活動や自然的現象を把握することができると考えた．例えば，ある地域のジオタグ付きマイクロブログの発言量の増減，または人々の移動の分析を行うことで，予測できなかった社会現象や地域的なイベントを発見することが可能であると考えられる．

そこで本研究では，お祭りや災害などのような地域イベントを発見するために，ジオタグ付マイクロブログを用いて通常とは異なる異常な活動が行われている地域を検知することを目的とする．そのため，地域ごとに通常のマイクロブログの発言傾向，人の存在傾向，人の移動量の傾向に基づきノーマルパターンを規定する手法を提案する．さらに，大量のジオタグ付マイクロブログを用いて異常活動の発生地域検知のための実験を行い，本手法の評価を行った．

本稿では，以下の構成で説明する．2 節で本研究の概要と関連研究について紹介する．3 節で異常活動の発生領域の検出手法を提案する．4 節で実験結果を示し，最後に 5 節でまとめと今後の計画について述べる．

2．マイクロブログによる地域の異常活動の検知

2.1　本研究の概要

本研究では，マイクロブログ分析による異常活動の発生地域の検知を行うことを目的とする．知を行う．

†兵庫県立大学大学院環境人間学研究科，Graduate School of Human Science and Environment, University of Hyogo

‡ 兵庫県立大学環境人間学部，　School of Human Science and Environment, University of Hyogo

そのため，大量のマイクロブログを取得し，地域ごとにマイクロブログからノーマルパターンを導き出し，それに基づく異常検以下に異常検知を行うための主な 4 つのステップを示すとともに研究の概略図を図 1 に示す．

i ）地域的マイクロブログ取得システムの開発
ii ）マイクロブログの発生分布による関心地域の設定
iii）通常の人々の振る舞い傾向の規定
iv）異常活動の発生領域の検知

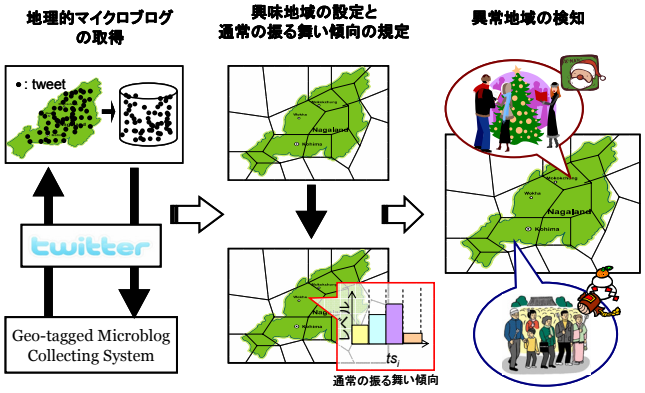

図1　本研究の概略図

まず，我々は地域に存在するより多くのマイクロブログを網羅的に取得するために，Twitter の API [6]を用いたマイクロブログ取得システムを開発した [3]．このシステムでは国レベルのような大きな地域から市や町のような狭い地域まで地域範囲の粒度をオンライン地図上で操作し指定することで，ユーザの様々な要求に合わせてその地域に存在するデータを取得することが可能である．

また，異常な活動をしている地域を検知する前処理として，分析対象となる地域をどのような基準で決定するか考える必要がある.ここでは，興味を持って分析する地域を関心地域と定義する．関心地域 (Region of Interests)を決定する簡単な手法として，その地域をある一定の大きさのセルに分割する手法や行政区域を用いる手法などが考えられる．

セルによる分割の場合，あらかじめセルのサイズをどのように適切に決定するかは困難である．例えば，ある大きさのセルに分割したとき，マイクロブログが一日 1 回しか挙げられないセルが隣接して存在する場合は，それらのセル一つ一つを分析することはあまり効率の良い分析とは言えず，セルを合わせて 1 つの領域と見なし分析すべきである．逆に,マイクロブログが多く存在するセルは，その中でも特に多く発信される場所が存在する可能性が高く，セル

のサイズよりももっと小さい単位のセルに分割すべきである．

次に，行政区域の場合だが，行政区域は国によって定められた領域であり，人が行動する活動地域は行政区域に沿って決まるものとは言いがたい．市レベルの行政区域にした場合，例えば姫路市にいる人は神戸市へ出勤に行く傾向が強くその領域を活動範囲としているにもかかわらず，それが考慮されない表面だけの領域区分けになる恐れがある．

そこで，関心地域の設定のため，我々は Twitter データの地理的分布による分析対象地域の分割手法を提案する．まず，取得システムから得られた大量のジオタグ付マイクロブログを用いて，ジオタグ（緯度・経度）に基づき K-means 手法 [2]より空間の分類を行う．これにより，実際の発言場所に基づく領域のセットが K 個得られる．本手法ではその K 個の関心地域を設定し，セットごとに得られたジオタグ中心点に基づき Voroni 図を作成する．これによって，領域の大きさを実データの発信傾向に合わせて分割することができるとともに，効率よく領域の異常検知分析が行えると考えられる．

しかし，異常検知を行うために，何が通常の状態で何が異常な状態か決める必要がある．そこで，まず我々は関心地域ごとに 1)通常どの程度の発言量がある領域か，2)どれくらいの人が存在する領域か，さらに 3)どれくらい人々の移動が活発に行われるのかを調査し，ある一定のタイムスロットごとノーマルな状態を定義する．例えば，1 日である領域を 4 つのタイムスロット（朝，昼，夕方，夜）で分析した場合を考える．この際，朝 (06:00～12:00)はマイクロブログの発言量が多く，昼 (12:00～18:00)と夕方 (18:00～24:00)はその数が減少し，夜 (00:00～06:00)になるとまた発信量が増加するっというような形で通常の人々の振る舞いを見つけることができる．本研究では以上の 3 つ尺度に基づく通常の人々の振る舞い（ノーマルパターン）を規定する．それによって，ノーマルな状態と比較したいある日の傾向を照らし合わせることで，異常な活動地域を検知することができる．

2.2　関連研究

人々の社会的な移動傾向を抽出する研究として，Wang [7]らは GPS が内蔵されたタクシーから移動軌跡のデータセットを分析し，人の流れを解明するために乗客がよく乗る場所，降りる場所といった"hot spot"の発見を提案した．この研究は，移動の傾向から重要な場所を検知する店に類似するが，データの収集手法で大きく異なる．この研究では，タクシー会社のサポートから多くのデータを集めるために時間とコストを要するが，我々はマイクロブログ取得システムを開発したことで，どんな領域でも容易に膨大なデータを取得することが可能である．

最近の研究として，Sasaki ら[4]は Twitter のリアルタム性に着目し地震や台風のような自然災害警告システムを開発した．この際，彼らはクエリとして自然災害に関するキーワードを用いて関連 tweet を収集しているが，探したい自然的イベントが変わるたびにクエリを変えなければならないことやクエリが実際に関連 tweet をどれくらい網羅できているか判断することが難しい．しかしながら，我々の研究は，クエリを用いず異常活動の発生地域をマイクロブログの発言量，人の存在量，人の移動量より様々な地域イベントの発見を目指す点で異なる．

3．異常活動の発生領域発見のためのノーマルパターン検出手法

異常な活動をしている領域検出モデルを図2に示す．領域ごとにマイクロブログの発言量，人の存在量，人の移動量の習慣的な傾向を，ジオタグ付きマイクロブログ基づき定義しノーマルパターンを規定する．次に，習慣的データ分布傾向からあるタイムスロットごとにそれぞれの領域でマイクロブログの発言量と人の存在量を比較し，ノーマルパターンとは異なる傾向を示す領域を抽出する．そして，その領域の移動量を通常時のノーマルパターンと比較することで，異常活動の発生領域を検知する．以下具体的な手法について述べる．

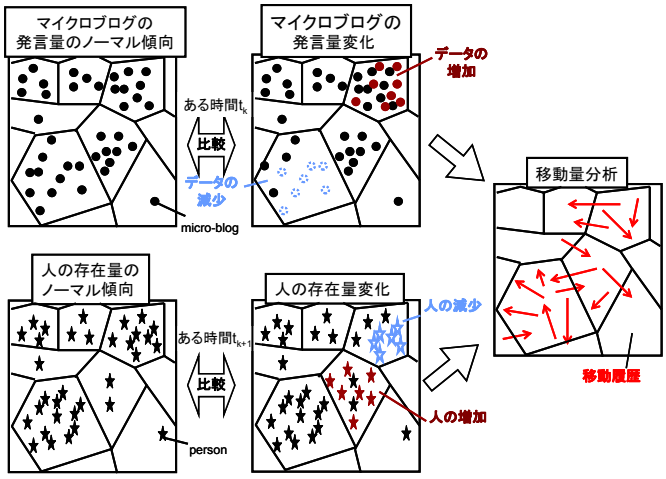

図2　異常な活動領域検出モデル

3.1　マイクロブログの発生分布による関心地域の設定

我々が開発したマイクロブログ取得システムからユーザ id によって下記のようにデータを分類する．

$$mb_i \in MB \qquad (1)$$

$$mb_{i.msg} = \langle \ldots\ldots, m_j, m_{j+1}, \ldots\ldots \rangle \qquad (2)$$

$$m_j = (t, loc, text) \qquad (3)$$

$$loc = (latitude, longitude) \qquad (4)$$

ここでMB はある対象地域において集めたすべてのマイクロブログデータを表す． $mb_{i.msg}$ は時系列にソートされたユーザの発信履歴を表す．ここでm_j は， j 番目のマイクロブログを表し，t時間に場所locからテキスト$text$を発言したことを表す．また，取得したMBは時間，位置（領域），テキストの3つ属性に基づき容易に抽出可能なデータ構造を構築した．例えば，取得した範囲内であれば，時間$[t, t_{k+1}]$にある領域から発言されたマイクロブログのみを抽出することが可能である．

そして，取得したジオタグ付きマイクロブログから緯度，

経度に基づき分析対象地域を領域分割する．異常活動が発生している領域を見つけるために，シンプルなクラスタリング手法として知られるK-means法によりマイクロブログの緯度，経度に基づきデータの分類を行う．これより，よくマイクロブログが挙げられる場所やそうでない場所など実際のマイクロブログの位置的な尺度に基づき分類できる．また，K-means法の場合，Kの数の決め方をどのように決めるかという問題が挙げられるが，今回実予備実験よりK=300に規定する．最後に，そのデータ集合の中心点からVoroni図を作成することで領域分割を行う．本手法で形成された領域集合Rを図3に示すとともに，この領域を固定して評価を行う．

$$R=\left\{ r_i \middle| \begin{array}{l} r_i \cap r_k = \phi \, and \, r_i, r_k \subset target_region, \\ where \, j,k \in Cluster_IDs \;\; and \;\; k \neq j \end{array} \right\} \quad (5)$$

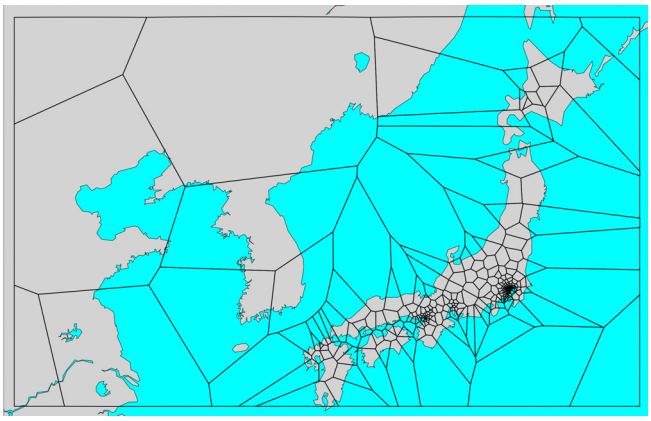

図3　K-mean + Voroni による日本地域における領域分割

3.2　ノーマルパターンの規定手法

ジオタグ付マイクロブログに基づき通常のマイクロブログ発言量，人の存在量，さらに人の移動量のパターンを測る．今回，それぞれの関心地域r_iによってデータが少ないもの，多いものなど様々であることが考えられたため，本研究では関心地域r_iごとにあるタイムスロット*ts*における3つの尺度から関心地域r_iのノーマルパターンを定義する．

最初に，以下のようにシステムから取得したデータを分類する．

$$MB\big|_{[s,e]} = \sum_{i=1}^{k} MB\big|_{[s+(i-1)ts,\, s+its)} \quad (6)$$

$$s: start\ time,\ e: end\ time,\ ts: time\ slot$$

これによって，タイムスロットごとの発言データを分類でき，スロット単位で評価を行うことができる．また，タイムスロットなどの粒度も 1 時間単位や一日単位など幅広く適応させることが可能である．

マイクロブログの発言量および人の存在量に基づくノーマルパターン決定手法について図4に沿って説明する．
1日以下のタイムスロットにおいてスロットごとに発言量を求め，数日分のデータより通常の発言傾向を定義する．まず，ある領域r_iの数日分のデータ<d_1，d_2，…，d_k>を用意し，それをタイムスロット*ts*によって分割する．本研究においては，morning(6:00-12:00), afternoon(12:00-18:00), evening(18:00-0:00), night(0:00-6:00)の4つのタイムススロットに分類する．次に，異なる日の同じタイムスロット同士に分類し，それぞれのタイムスロットにおけるノーマルな発言レベルを決定する．

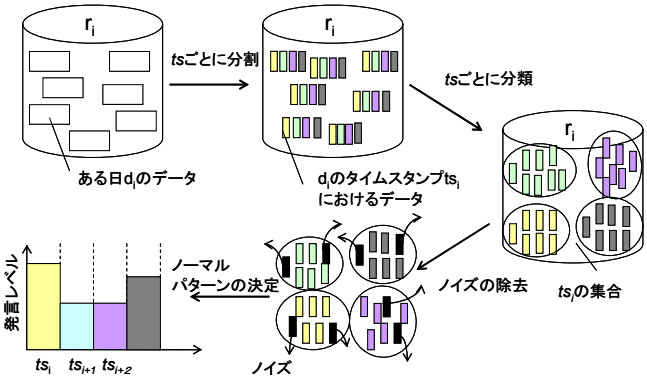

図 4　発言量に基づくノーマルパターン決定手法

この際，単純にそれぞれの平均をノーマルとするとデータのばらつきなど考慮できないため，集合内の最小値と最大値からノーマルな範囲を定める．しかし，ノーマルを定めるために用いたデータ内でも異常なデータが含まれている可能性が考えられた．そこで，本研究では5点要約法による箱髭図 (boxplot) [1]を用いてノーマルな範囲を決定する．

移動量に基づくノーマルパターン決定手法について説明する．r_iにおける移動量とは，あるタイムスロット*ts*においてr_i内に存在するユーザがある期間においてどれくらい移動距離を持っているかを表す．ある期間とは，タイムスロット*ts*よりも長い期間を表す．移動の種類として3つを定義している．

i) r_i内での移動 (inner movement)
ii) r_i内に入ってくる移動 (incoming movement)
iii) r_iから出て行く移動 (outgoing movement)

今回この3つの移動タイプの起こった回数の合計を移動量とし，図3で示した発言量および存在量と同様な流れでr_iの移動量に基づくノーマルパターン決定を行う．

3.3　異常活動の発生地域の検知手法

異常活動の発生領域を検知するために発言量，存在量，移動量をノーマルパターンと比較するが，どれくらい発言量や存在量が増加および減少したら抽出するかという閾値の問題が挙げられる．例えば，ノーマルな範囲を100～200とした場合，210は確かに増加したと言えるが，ノーマル範囲から見た時に顕著に増加したとは言いにくい．そこで，本研究では異常値を下記の手法で決定する．

始めに，ノーマル範囲をそれぞれの領域ごとに箱髭図により作成しているため，それに基づき異常範囲を決定する．図5のように，ボックスの中は安定領域とし，点線で示された髭のように生えている部分はノーマル範囲を作成する際に外れ値を除いた最大値と最小値を表しているため，許容範囲といえる．そこで，本研究ではそれ以外の領域を異常な領域とし評価を行う．これにより，異常値の上限を超

える値または下限を下回る値の場合，それに該当する領域r_iを抽出する．

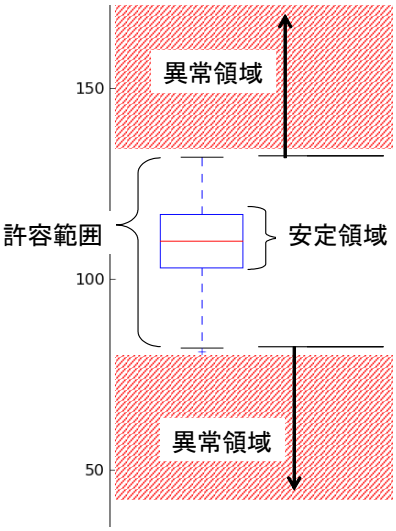

図5　箱髭図を用いた異常範囲の決定

そして，3つの尺度を組み合わせから異常活動の発生地域とする組み合わせを決定する．以下にすべての組み合わせの表を示す．

表1　異常検知尺度の組み合わせによる最終状態

	発言量 (M)	存在量 (P)	移動量 (D)	最終状態 (F)
(a)	N	N	N	N
(b)			A	N
(c)		A	N	N
(d)			A	A
(e)	A	N	N	N
(f)			A	A
(g)		A	N	N
(h)			A	A

N: normal　　A: abnormal

表1に示したように，(a)から(h)までの8つの組み合わせが考えられたが3の尺度のうち2つ以上異常値（abnormal)であり，かつ移動量が異常値を示すものを最終状態として異常な状態であるとした．これについて，最終的に求まる異常検知式を論理関数で表すと下記のようになる．

$$F = D'(MP' + M') \qquad (7)$$

式（7）において，通常状態をM, P, D，異常状態をM', P', D'とすると，最終状態Fは移動量が常に異常状態であることが最初に言え，かつ発言量が通常状態，存在量が異常状態であるか，または発言量が異常状態であるときとなる．表1の場合では，(d), (f), (h)が異常な状態となる．それぞれについて図を用いて説明する．

例えば，(h)の場合，図5のようにノーマルな状態においてマイクロブログの発言量および人の存在量が少ない領域r_iが，ある時間t_kにデータの増加が見られる．この際に，人の移動量を分析しても，移動量の増加が見られる場合，その領域は異常活動の発生領域であるとし，検知を行う．

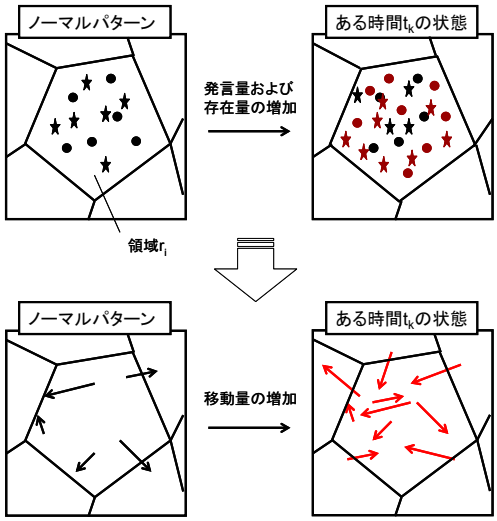

図6　(h)に基づく異常活動の発生地域検知の一例

また，(f)について図6を用いて説明する．マイクロブログの発言量がノーマルパターンと比べてある時間t_kに増加または減少の異常が見られるとき，人の移動量を測定すると，移動量の増加または減少の異常が見られる．この時，その領域は異常活動の発生領域であるとし，検知を行う．

図7の場合において，移動量の増加の場合，例えば都市部でのお祭りの際に通常よりも多くの人々がその地域に集まることが考えられる．一方，移動量の減少の場合，例えば都市部での電車のトラブルがあり運転を見合わせるときに，通常よりも話題性としてデータ量は多くなるが移動量は電車の移動がない分だけ通常よりも減少することが考えられる．

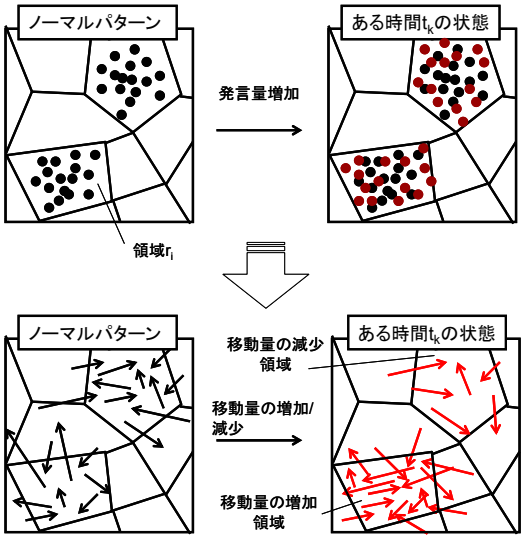

図7　(h)に基づく異常活動の発生地域検知

最後に，(d)について図8を用いて説明する．ノーマルパターンと比較して，人の存在傾向に増加または減少の異常が見られるとき，人の移動量を測定すると，移動量の増加または減少の異常が見られる．この時，その領域は異常活動の発生領域であるとし，検知を行う．

移動量の増加の場合，例えば夏のお盆の際に都市部で活動している人々は一斉に実家の方に帰省するため，その場所でも発言量は減少するが移動量としては普段より遠くへ行く移動が多くなり，増加することが考えられる．一方，移動量の減少の場合，神戸で起こったインフルエンザの時のように普段は出勤や通学のために多くの人が行き交う場所が，会社や学校の臨時の休みによって人が少なくなり，それに伴い移動量も減少することが考えられる．

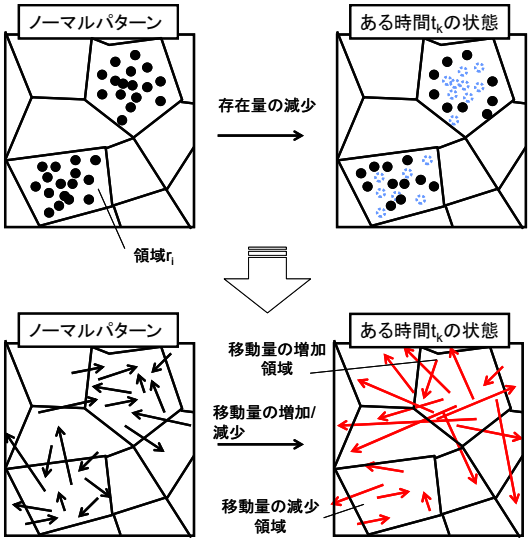

図8　(d)に基づく異常活動の発生地域検知の一例

4.　実験と評価

本実験において，提案した地域イベント発見のための異常活動の発生領域を検知できるか検証を行った．また，地域イベントとしてはお祭りなどの人が集まるイベントを検出したいと考えたため，人の存在量が増加しかつ移動量も増加した領域を抽出しその評価を行った．

4.1　データセットと関心地域の設定

地域的マイクロブログ取得システムを用いて日本領域（latitude range : [30.004609; 45.767523], longitude range : [116.27921; 148.381348])においてジオタグ付マイクロブログを約 1 ヶ月半分(2010/06/04-2010/07/24)取得した． その結果，366,556 のマイクロブログユーザから 21,623,947 のデータを取得できた．

この取得データから，関心地域の作成を行った．取得したデータすべてを用いて 3.1 章でも述べたように K-means クラスタリング手法に基づく空間の分割を行った．そして，実空間に存在するデータに基づき 300 個の関心地域を設定した．図 3 にその結果を示しているが，領域のサイズが大きい地域と小さい地域が存在することが分かる．
これは，データの存在傾向を表し，マイクロブログが多く存在する都市部では小さな関心地域が多く見られ，マイクロブログが少ししか存在しないときは，比較的大きな関心地域となっていることが分かる．

4.2　関心領域ごとのノーマルパターンの設定

本実験では，6 月 8 日から 6 月 30 日の約 3 週間のデータから，それぞれのタイムスロットごと（morning, afternoon, evening, night)に各領域 r_i の人の存在量と移動量に基づくノーマルパターン規定した．その結果の 1 つとして，京都の鞍馬口駅周辺の人の存在量に基づくノーマルパターンを図 9 に示す．図 9 のようにタイムスロットごとに，人の存在量，人の移動量に基づきノーマル範囲を規定した．今回は領域 R=300 個のノーマルパターンをそれぞれで作成したが，図 9 にも見られるように night，morning，afternoon，evening の順に発言量が多くなる傾向が多くの領域で見られた．

本実験では，先述したように箱髭図の box から伸びる点線と交わる横線の最大値，最小値を超える値を異常値とした．例えば，図 9 のタイムスロット morning の場合，最大値 62 を超えた場合，異常となる．

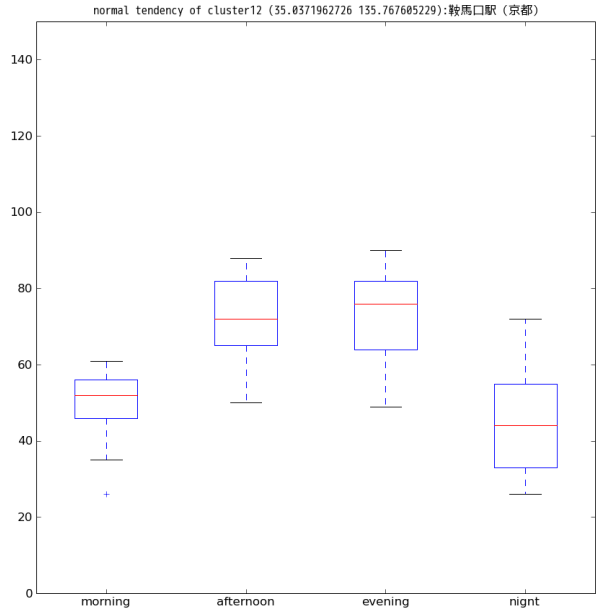

図 9　京都の鞍馬口駅周辺の人の存在量のノーマルパターン

4.3　異常活動の発生地域の検知

ノーマルパターンで用いていない 7 月 10 日から 7 月 20 日の 10 日間のデータを用いて 1 日ごとに 4 つのタイムスロットと比較して異常な活動が行われている領域 r_i を検知した．本実験においては，先に述べたようにまず人の存在量の比較を行い，そこで異常値を示した場合さらに人の移動量の比較を行う．これによって，導き出された領域を異常活動の発生地域として抽出した．

領域の数が 300 個, タイムスロットが 4 つ, 日数が 10 日であるので, 比較として 12,000 回行った．その中で異常と判断されたタイムスロットの合計は 749（morning:239, afternoon:298, evening:111, night:101)であった．つまり，全体の約 6%が異常活動の発生している領域であると判定された. 図 8 で示した地域においても 7 月 17 日の afternoon と evening で図 10 のように異常値を示した．

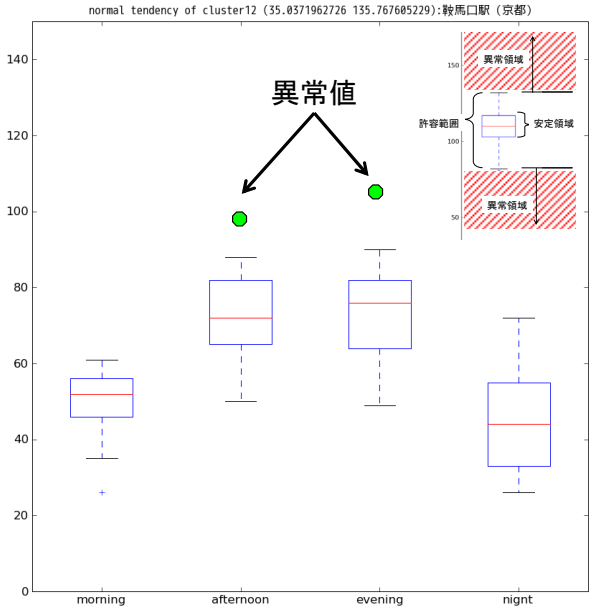

図 10　京都の鞍馬口駅周辺の人の存在量の
ノーマルパターンと 7 月 17 日の傾向比較

そこで，異常な活動が行われていると判定された領域のいくつかを取り上げ，領域内のメッセージの分析を行った．今回は，全体的に 7 月 17 日に異常を示す領域が多く，その日を中心に分析を行ったところ，図 9，10 で示した京都の鞍馬口周辺において祇園祭についてのマイクロブログが見られた．実際に，図 11 にように京都の鞍馬口駅周辺から挙げられた祇園祭に関するマイクロブログを示す．この際に，メッセージだけでなく写真やビデオまでも挙げられていた．

図 11　京都の鞍馬口駅周辺から挙げられた
マイクロブログ

その他にも，愛知の領域でも名前は同じだがその地域で行われていた祇園祭について書かれているマイクロブログが多く，また花火大会が行われていた地域が検知されていた．

提案した手法により異常活動の発生地域の検知を行い，お祭りのような地域イベントを発見することができた．しかし，抽出された中には，なぜ異常領域と判定されたか判断が難しいものもあり，それをどのように分析していく今後の課題である．また，今回は人が集まるような活動の検知を行ったが，マイクロブログが少なくなった領域などの検証も今後行う予定である．

5.　まとめと今後の課題

本研究では，ジオタグ付マイクロブログを用いて異常活動の発生領域検知を行い，お祭りのような地域イベントの発見手法を提案した．大量のマイクロブログ取得システムを開発し，そこで取得したデータから実データの発信傾向に合わせて領域分割を行う独自の分析対象地域の分割を行った．そして，領域ごとに通常の状態を表すノーマルパターンを定義し，ある時間における状態と比較することで異常活動の発生領域を検知する手法を提案した．

今後の課題としては，さらなる地域イベント発見のための手法の検討が挙げられる．発言量，存在量，移動量の 3 つの尺度を提案しているが，それをどのように組み合わせて分析するか考察する．また，移動の種類として 3 つ定義したが，それぞれ分けて分析を行うことで移動の種類の比較による異常活動の検知を行うとともに，移動によるイベントのタイプ分類も行う予定である．

6.　謝辞

本研究の一部は，平成22 年度科研費基盤研究(B)(2)「ユーザの潜在的意図を用いたレス・コンシャス情報検索基盤の構築」(課題番号：20300039) および平成22 年度特別研究員奨励費(21.197) によるものです．ここに記して謝意を表すものとします．

7.　参考文献

[1] boxplot：
http://bstat.f7.ems.okayama-u.ac.jp/statedu/lispstat-book/node30.html
http://aoki2.si.gunma-u.ac.jp/lecture/Dosuu/box-and-whisker.html
http://matplotlib.sourceforge.net/api/pyplot_api.html#matplotlib.pyplot.boxplot

[2] K-Means: Lloyd., S.P., Least squares quantization in PCM, IEEE Transactions on Information Theory 28 (2), pp. 129-137

[3] Tatsuya Fujisaka, Ryong Lee and Kazutoshi Sumiya Discovery of User Behavior Patterns from Geo-tagged Micro-blogs, 4th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2010), pp.246-255, Suwon, Korea, January 14 2010

[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, Proc. 18th International World Wide Web Conference (WWW2010), April 2010

[5] Twitter: http://twitter.com/

[6] Twitter Open API: http://apiwiki.twitter.com/

[7] Handong Wang, Haixiang Zou, Yang Yue, and Qingquan Li. Visualizing hot spot analysis result based on mashup. Proceedings of the 2009 International Workshop on Location Based Social Networks, 2009