2B4-NFC-02a-1

ソーシャルメディアを用いた局所地域における異常検知手法の提案

Anomaly Detection Local Areas Using Social Media

榊 剛史 *1 Takeshi Sakaki　松尾 豊 *1 Yutaka Matsuo　鳥海 不二夫 *1 Fujio Toriumi　篠田 孝祐 *2 Kosuke Shinoda　栗原 聡 *3 Satoshi Kurihara　風間 一洋 *4 Kazuhiro Kazama
野田 五十樹 *5 Itsuki Noda

*1東京大学 The University of Tokyo　*2慶応義塾大学 / 理化学研究所 Keio University / RIKEN　*3電気通信大学 The University of Electro-Communications
*4和歌山大学 Wakayama University　*5産業技術総合研究所 Advanced Industrial Science and Technology

In an emergency situation, identifying devastated areas is crucially important because it needs for safety confirmation, display of evacuation routes, and making plan for support disaster victims. There are some infrastructures for this purpose. However, those infrastructures possible do not work correctly when a disaster strikes. While, a vast amount of social media data is created sequentially with the growing popularity of social media. This paper treated such social media data as a kind of data streams and try to identify devastated areas by applying existing anomaly detection methods for data streams.

1. はじめに

災害発生時に，いち早く災害の発生地域を検知し被災地域を特定することは，安否確認や避難経路決定，災害支援対策立案等の観点から必要不可欠な技術であり，そのための様々なインフラが設置されている．しかし，災害発生時にはそれらが機能しない可能性もある．実際，東日本大地震においては，災害地域を把握し，政府に伝達すべき地方自治体が地震によって機能停止し，正確な被害状況の把握が遅れ，そのために災害復旧の初動が遅れたとされている *1．

一方，ソーシャルメディアの普及に伴い，人々の日々の活動や現在の状態，様々なものごとに対する意見や感想などがソーシャルメディア上に数多く投稿されるようになった [Naa 11]．つまり，ソーシャルメディア上にはユーザが観測した事柄やユーザ自身の思想，意見に関するデータが大量に存在している．Twitter や Facebook を初めとするソーシャルメディアは大規模なユーザを抱えているため，ユーザの行動データが途切れること無く常に蓄積され続けている．これより，大規模ソーシャルメディア上のデータは，一種のストリームデータと考えられる．本稿では，このような大規模ソーシャルメディア上のユーザ行動記録の蓄積をユーザ行動ストリームと呼ぶ．

また，近年はビックデータという言葉がトレンドトピックとして注目されている．近年は，電子化された大規模なデータが企業内外に集積され，またそれらのデータを分析するためのツール及びハードウェアが整備されている [岩爪 13]．そのような状況において，それらの大規模データを分析し，そこから得られる知見を活用することで，企業経営や製品開発をより戦略的かつ効率的に行おう，という考えが徐々に浸透し始めている．このような傾向の中で，異常検知の手法が注目を集めている．例えば，ネットワーク監視というタスクを考えて場合,「興味があるのは正常時の振る舞いというよりはむしろまれに生ずる何かの異常な状況で，全体の分布のすそ（tail）の部分に存在する現象が重要」であると考えられる [武田 12]．このような「通常とは違う振る舞い」の発見はリスク管理という観点から，様々な業種，タスクにおいて重要になると考えられる．

ソーシャルメディアには，人々が自分の周囲の出来事について投稿するため，災害発生時には，なんからの異常が ユーザ行動ストリームに現れると考えられる．ここで，ビックデータにおける異常検知の考え方を災害発生時のソーシャルメディアに適用すると，被災地域のユーザによるユーザ行動ストリームには何らかの以上が発生すると推測される．つまり，ある地域におけるユーザ行動ストリームで異常が検知された場合，その地域が被災地域であると考えられる．

本稿では，ソーシャルメディア上のユーザによる投稿をユーザ行動のストリームデータ捉え，それに異常検知の手法を適用することで，実際に災害を含めた何らかの異常が発生している地域を検出することを目指す．具体的には，各地域ごとにその地域内で行われたユーザ行動ストリームに異常検知の手法を適用し，異常が検知された地域を被災地であるとみなす．

データとしては東日本大震災の Twitter データを利用する．また，取り扱うユーザ行動としては，投稿行動及び 2 種類のユーザインタラクションを分析する．

2. 関連研究

いくつかの研究において，災害時にどのように Twitter が活用されるかの分析が行われている [Aramaki 11]．例えば，Mendoza らは 2010 年のチリ地震において，地震後数時間後から数日後まで，Twitter ユーザがどのように振舞ったかについて調査を行ない，その特徴を明らかにしている [Mendoza 10]．これらの研究では，災害発生時にはソーシャルメディア，特に Twitter がよく活用されることが言及されている．

異常検知について，異常検知は大規模データの中から異常なデータや変化を検出する技術であり，ネットワークトラフィックの障害検知や不正アクセスの早期発見などに用いられてい

連絡先: 榊 剛史，東京大学大学院工学系研究科，東京都文京区弥生 2-11-16 工学部 9 号館

*1 http://www.keieiken.co.jp/monthly/2011/1109-05/index.html

る [Mislove 07]．異常検知の手法は，大きく分けて 3 つに分類できる．多次元ベクトルにおける外れ値を検出する「外れ値検出」，多次元時系列データから急激な変化を検出する「変化点検出」，多数の状態遷移データから異常行動を検出する「異常行動検出」である [山西 09]．

本研究では，時系列データを扱うため，変化点検出の手法が必要となる．その中でも汎用的な手法として認知されている Change Finder を適用する．

3. データセット

本節では，本研究で用いたデータセットについて述べる

3.1 データセット

本研究では，震災前後に投稿された日本語のツイートを収集した．収集手順は，以下の通りである．

1. 日本語 Twitter ユーザのリストを作成
2. 1. のリストのユーザが震災前後に投稿したツイートを収集

収集したデータセットの詳細は以下の通り．

- ユーザリストに含まれるユーザ数: 130 万ユーザ
- ツイート数:356,118,522 ツイート
- 収集期間: 3 月 7 日〜3 月 24 日

3.2 ユーザ居住地の推定

本研究では，地域別にユーザ行動ストリームを収集するために，各ユーザの居住地推定を行った．

Twitter のユーザプロフィールにはユーザの居住地を入力する項目がある．ただし，ユーザ毎に正確性や具体性が異なる．例えば,「東京都北区王子」のように大字まで入力しているユーザもいれば,「東京都」や「日本」のように大ざっぱな記入をしているユーザ，さらには『夢の中」「この世のどこか」など実際の居住地とは無関係の情報を入力しているユーザもいる．そこで，本研究では推定可能なユーザのみ居住地を推定し，それを場所情報として利用した．ユーザの居住地は，市町村名及び都道府県名（漢字，ひらがな，カタカナ，アルファベット表記いずれか）がユーザプロフィールに含まれているか否かにより判定した．全 130 万ユーザに対し，425,178 ユーザについて都道府県を判定することができた．

3.3 代表的な都道府県の選定

本稿においては，47 都道府県すべてを分析対象としている．ただし，紙面の都合上，すべての都道府県についての結果を示すことは困難である．そこで，東日本大地震において代表的な特徴を持つ都道府県を選定し，それらについての異常検知結果を示すものとする．

選定した都道府県は表 1 の通り．被害の大きかった地域として岩手県，宮城県を，被害が中程度だがそれ以外の影響の大きい事象が起きた地域として，福島県，東京都を，被害のほぼなかった地域として大阪府をそれぞれ選定した．

4. 異常検知手法

本節における課題は，ユーザ行動ストリームデータからの異常検知であるため，時系列データから異常検知を行う変化点検出の手法が適用可能であると考えられる．

変化点検出の手法は数多く提案されているが，本研究では汎用的な手法の一つである *Change Finder* を適用する [山西 09]．Change Finder は忘却型学習アルゴリズムを採用した変化点検出エンジンであり，非定常な時系列データに対し，既存手法より少ない計算量で，変化点を検出できる手法である．

変化点検出の基本的な考え方は，過去の時系列データから将来の値を予測し，その予測値と観測値に乖離が生じた点を変化点として検出する，というものである．従来の変化点検出においては，時系列データの予測には AR(Auto Regressive) モデルやその派生である ARMA(Auto Regressive Mean Average) モデルが用いられてきた．AR モデルとは，ある時系列データにおいて定常状態を仮定し，時系列データの過去の変化から定常状態を予測するモデルである．しかし，定常であることを仮定しているため，非定常的なデータには適用困難である．また，バッチ学習方式であるため計算量が多く，リアルタイムな異常検知には適していない．

それに対し，SDAR(Sequentially Discounting Auto Regressive) モデルはオンライン忘却型学習アルゴリズムである．忘却型アルゴリズムであり過去の統計量の影響を減らすことができるため，非定常的なデータにも適用可能である．また，オンライン型アルゴリズムであり，過去のパラメータと直前の時系列データのみで予測が可能であるため，計算量が少なく，リアルタイムに予測することが可能となる．

$P\left(x_i, |x^{i-1}, \theta\right)$ を入力時系列データ x_i に対する確率密度関数とすると，SDAR アルゴリズムは下記の様な尤度 I を最大化するようにパラメータ θ を推定する手法である．r は忘却パラメータであり，i 時点前の影響を $(1-r)^i$ 倍に減少させている．

$$I = \sum_{i=1}^{t} (1-r)^i \, logP\left(x_i, |x^{i-1}, \theta\right)$$

Change Finder のアルゴリズムは下記の通り．

1. SDAR アルゴリズムにより入力時系列データの確率密度関数を学習
2. 各時点での値について学習した確率密度からの外れ程度によりスコアリング
3. 各スコアの移動平均を取ることで平滑化
4. SDAR アルゴリズムにより平滑後スコアの確率密度関数を学習
5. 各時点での平滑後スコアについて，学習した確率密度から外れ程度によりスコアリング

本稿では，この Change Finder をデータセットから得られる各地域のユーザ行動ストリームデータに適用することで，異常が発生している地域を検出することを目指す．

5. ユーザ行動ストリームにおける異常検知

本節では，Twitter ユーザの行動ストリームデータに Chage Finder を適用することで，被災地域が特定できるかどうかの検証を行う．

表 1: 代表的な 5 都道府県

都道府県	被害	人口	その他
岩手	壊滅的	中程度	-
宮城	壊滅的	中程度	-
福島	大きい	中程度	原子力発電所事故
東京	中程度	大規模	計画停電/帰宅困難者
大阪	無し	大規模	-

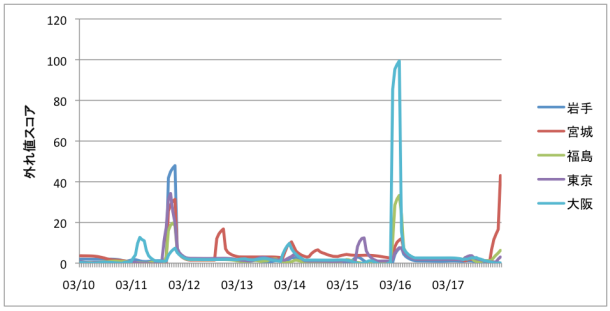

図 1: 5 都府県における投稿数変動による外れ値推移

データセットからは様々な Twitter ユーザの行動ストリームデータを取得できるが，本研究では，下記の 2 種類のデータを用いる．

投稿数 各ユーザにより投稿されるツイート数

インタラクション数 2 ユーザ間で行われるインタラクションの回数．Reply と RT と 2 種類のインタラクションが考えられる．

榊らの分析により，これら 3 種類の行動ストリームデータは，震災前と震災直後で大きく変化した事が明らかにされている [Sakaki 13]．

また，インタラクション数では，それぞれのユーザが情報発信者と情報受信者に分かれる．そのため，各地域のユーザ行動ストリームにおいて，各地域の情報発信数変動における異常検知，情報受信数変動における異常検知に分けて異常検知を行う．

さらに本節では，投稿数とインタラクション数を組み合わせた指標の変動からも異常検知を行う．

5.1 投稿数変動における異常検知

震災前後における各地域の投稿数の変動データに対して，Change Finder を適用し，異常が発生した時間及地域の検出を行う．

今回の実験では，学習のウィンドウ幅を 24 時間，平滑化のウィンドウ幅を 10 時間とした．投稿数は 1 時間単位で集計して取り扱うものとする．また紙面の都合上，3 月 10 日〜17 日の 8 日分のデータについて表示する．

図 5.1 に代表 5 都道府県の外れ値スコアのグラフを示す．図 5.1 より，どの都道府県でも東日本大地震発生直後の 3 月 11 日 15 時〜17 時付近及び東海沖を震源とする震度 6 の地震が発生した 3 月 15 日 10 時〜3 月 16 日 1 時付近で大きな値を示している．これより，単純に外れ値スコアが大きく変動した時刻・地域を手掛かりに，被災地を推定することは困難であると考えられる．

そこで，外れ値スコアが大きく変動した時刻における各地域における外れ値の大きさを比較する．各地域の外れ値グラフにおいて，同時に最大値を記録した地域数が最も多かったのが 3 月 11 日 17 時の時点である．この時点における各地域の外れ値スコアにおいて，上位 5 都道府県を表 5.2 に示す．

被害の大きかった岩手県が 3 位にランクされている．また 1 位の山形県，2 位の北海道は比較的震源地に近い地域である．しかし，本来であれば震源地に最も近い宮城県及び岩手県が 1 位，2 位にランクされるべきであり，また，4 位，5 位は震源地から遠く離れた地域である．このような結果となる理由としては，例えば岩手県内でも無事であったユーザは地震に喚起されて投稿数が増え，他方被害の大きい地域にいたユーザは投稿数が 0 になり，結果として投稿数の被災による影響が相殺されてしまっている可能性が考えられる．また，震源地から遠く離れた地域のユーザは問題無く投稿できるため，地震に喚起されて投稿数が増え，結果として大きな外れ値を記録した可能性が考えられる．

表 2: 投稿数/インタラクション数変動の外れ値上位 5 都道府県 (3 月 11 日 17 時)

順位	投稿数	Reply		Retweet	
		送信	受信	送信	受信
1	山形	神奈川	茨城	宮崎	山梨
2	北海道	東京	千葉	大阪	熊本
3	岩手	山梨	新潟	愛知	埼玉
4	山梨	新潟	山梨	愛媛	三重
5	兵庫	沖縄	埼玉	埼玉	千葉

いずれにせよ，この結果から考えて，各地域の投稿数変動から被災地推定を行うことは困難であると考えられる．

5.2 インタラクション数変動における異常検知

震災前後における各地域間のインタラクション数の変動データに対して，Change Finder を適用し，異常が発生した時間及地域の検出を行う．

ここでは Twitter におけるインタラクション機能である Reply，Retweet をユーザインタラクションとして用いる．また，各地域におけるインタラクションを情報発信数の変動，情報受信数の変動に分けて考える．今回の実験では，学習のウィンドウ幅を 24 時間，平滑化のウィンドウ幅を 10 時間とした．インタラクション数は 1 時間単位で集計して取り扱うものとする．また紙面の都合上，3 月 10 日〜17 日の 8 日分のデータについて表示する．

本実験においても，投稿数変動と同様に，3 月 11 日 15 時〜17 時付近及び月 15 日 10 時〜3 月 16 日 1 時付近で殆どの地域が大きな外れ値を記録している（投稿数変動と同じ傾向であるため，グラフは省略する）．これより，インタラクション数変動においても，単純に外れ値スコアが大きく変動した時刻・地域を手掛かりに，被災地を推定することは困難であると考えられる．

そこで，ここでも同様に 3 月 11 日 17 時における各地域の外れ値スコアにおいて，上位 5 都道府県を表 5.2 に示す．

表 5.2 より，Reply は関東南西側，つまり関東内で震源地から遠い地域かそれより震源地より遠方で急激に発信されるようになったことが分かる．Reply の受け手としては関東の北東側の地域が上位にランクされている．Retweet に関しては，送受信共に震源地より遠方で急激に変化したことが伺える．

この結果から考えて，Reply はある程度震源地からの距離を反映していと考えられる．しかし，各地域のインタラクション数変動から被災地推定を行うことは困難であると考えられる．

5.3 外れ値スコアの変化傾向を考慮した被災地推定

投稿数変動，インタラクション数変動，いずれの異常を検知しても被災地を推定することが困難であることがこれまでの実験でわかった．これは，様々な理由が考えられるが，最も大きな理由は，大規模な災害発生時は投稿数，Reply 数，Retweet 数共に全国的に増加してしまうために，どの地域でも異常が検知されてしまうために，それぞれ単体では被害の程度を判別することは難しい．

ここで，実際に被害の大きい地域において，ソーシャルメディア上のユーザ行動にどのような変化が発生するかの仮説を検討してみたい．まず，全国的にソーシャルメディアの活動

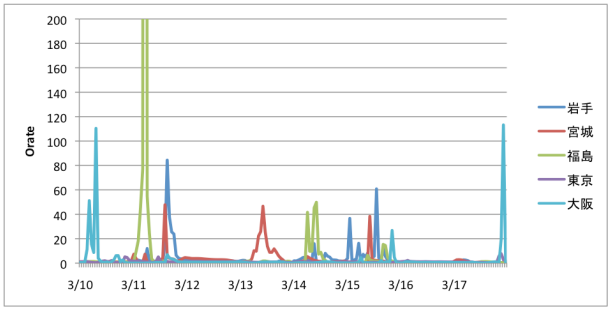

図 2: 5 都府県における O_{rate} の推移

表 3: O_{rate} 上位 5 都道府県 (3 月 11 日 15 時)

順位	都道府県名	O_{rate}
1	宮城	154.34
2	岩手	62.20
3	群馬	14.52
4	沖縄	14.28
5	長野	13.11

が活発になるため，被災地域が受け取るインタラクション数は増加すると考えられる．すなわち，インタラクションの受信地域としては外れ値が大きくなると考えられる．しかし，被災地域ではソーシャルメディア上で何らかの行動を起こすことが困難になるため，投稿数は減少すると考えられる．また送り出すインタラクション数は大きく変化しないと考えられる．すなわち，投稿数の外れ値は大きくなり，またインタラクションの発信地域としては外れ値が小さくなると考えられる．ここで，地域 l のある時点 i での受信 Reply の外れ値スコアを $O_{i,l,rec}$，送信 Reply の外れ値スコアを $O_{i,l,send}$ であるとき，下記のような地域 l, 時点 i でのスコア $O_{ratei,l}$ を考える．

$$O_{ratei,l} = \frac{O_{i,l,rec}}{O_{i,l,send}}$$

被災地域においては分子でありの受信 Reply の外れ値スコアが増加し，分母である送信 Reply の外れ値スコアが減少するため，$O_{ratei,l}$ の値が大きくなると仮定できる．

そこで，実際に各地域での O_{rate} を算出する．図 5.3 は，代表 5 都道府県の O_{rate} の変動を表している．図 5.3 より，震災直後，岩手県，宮城県の O_{rate} の値が大きくなっている．また，福島原子力発電所事故が注目を浴び始めた 14 日以降は福島県での O_{rate} の値が大きくなっている．

また表 5.3 に，3 月 11 日 15 時において O_{rate} の上位 5 都道府県，及び O_{rate} の値を示す．全都道府県の中でも深刻な被害を被った岩手県，宮城県のスコアがその他の都道府県の値と比べて 5〜10 倍程度の値を示している．

この結果は，被害の大きい地域では O_{rate} の値が大きくなる，という仮説が正しい可能性があることを示している．ただし，この一例だけでは不十分であり，他のデータセットを用いてさらに検証をしていく必要がある．

6. 終わりに

本稿では，ソーシャルメディアから得られる地域毎のユーザ行動ストリームに異常検知の手法を適用することで，災害が発生している時間及び地域の検出を試みた．データとしては，東日本大地震における Twitter のデータを利用した．

まず，各地域の投稿数変動やユーザインタラクション数変動に高速かつ非定常データに適用可能な手法により異常検知を行った．その結果，個別のユーザ行動ストリームでは，被災地を推定することが困難であることが分かった．

次に，被災地において，各ユーザ行動ストリームの外れ値がどのように変化するかの仮説を構築し，それに基づいて，被災地らしさを表すスコアを提案した．そして，各地域毎にこのスコアを算出し，実際の被害の程度と比較を行った．この論文で扱うデータについては，被害の程度が深刻であった地域において提案スコアが高くなることが検証できた．

ただし，1 つのデータでは提案スコアの有効性検証には不十分である．今後様々なデータを用いて，提案スコアの有効性を検証すると共に，ソーシャルメディア上のユーザ行動ストリームデータに対して，既存の異常検知手法が有効であることを示していきたい．

7. 謝辞

本研究を行なうにあたり，ツイートデータの収集に協力していただいたクックパッド株式会社の兼山元太氏及び株式会社ホットリンクに感謝する．また，本研究は Microsoft Research Asia University Relations の助成を受けた．

参考文献

[Aramaki 11] Aramaki, E., Masukawa, S., and Morita, M.: Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter, in *Proceedings of the 2011 EMNLP*, pp. 1568–1576 (2011)

[Mendoza 10] Mendoza, M., Poblete, B., and Castillo, C.: Twitter under crisis: can we trust what we RT?, in *Proceedings of the SOMA 2010*, pp. 71–79, New York, New York, USA (2010), ACM Press

[Mislove 07] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B.: Measurement and analysis of online social networks, *Proceedings of the 7th ACM SIGCOMM IMC 2007*, Vol. 40, No. 6, p. 29 (2007)

[Naa 11] Hip and Trendy : Characterizing Emerging Trends on Twitter, *Journal of the American Society for Information Science and Technology*, Vol. 62, No. 5, pp. 902–918 (2011)

[Sakaki 13] Sakaki, T., Toriumi, F., Shinoda, K., Kazama, K., Kurihara, S., and Noda, M. Y., Itsuki: Regional Analysis of User Interactions on Social Media in Times of Disaster, in *Proceeding of the WWW 2013 Poster Session*, ACM (2013)

[岩爪 13] 岩爪道昭:特集「ビッグデータと AI」にあたって, 人工知能学会, Vol. 28, No. 1, pp. 82–83 (2013)

[山西 09] 山西健司:データマイニングによる異常検知, 共立出版 (2009)

[武田 12] 武田浩一, 井手剛:ビッグデータ処理の展望, in *PROVISION Winter 2012 No.72*, No. 72, 日本アイ・ビー・エム (2012)