1C3-OS-12-2

東日本大震災時のTwitterデータを用いた単語間の関係の時系列変化の分析

Time-Series Variation Analysis of Relationships between Terms Using Tweets in East Japan Earthquake

風間 一洋[*1] Kazuhiro KAZAMA　鳥海 不二夫[*2] Fujio Toriumi　榊 剛史[*2] Takeshi Sakaki　篠田 孝祐[*3] Kosuke Shinoda　栗原 聡[*4] Satoshi Kurihara　野田 五十樹[*5] Itsuki Noda

[*1]日本電信電話株式会社 Nippon Telegraph and Telephone Corporation
[*2]東京大学 The University of Tokyo
[*3]理化学研究所 Riken
[*4]大阪大学 Osaka University
[*5]産業技術総合研究所 The National Institute of Advanced Industrial Science and Technology

As a result of synchronization and concatenation of messages on Social Media, related terms for a big event such as earthquakes and football games have similar histograms of term frequencies. This paper presents a method, which repeat measuring the the earth mover's distance (EMD) between their sub-histograms within a window and sliding the window, to analyze time-series variation of relationships between such terms, We extracted nouns, which are frequently used in the East Japan Earthquake, from three hundred million tweets and analyze time-series variation of relationships between them. The results shows the relationships between terms change according to real events and opinions.

1. はじめに

東日本大震災において Twitter は震災発生時の情報交換手段と震災後の情報拡散手段として活用された [総務省 11]．このような大きな事件においては，フォローネットワーク上の情報の伝播の同期や連鎖が顕著に発生し，その結果として，その事件に関係する単語の出現頻度の時間的変化は時間軸に対して類似したパターンを生じる．そこで我々は以前に Twitter のツイートに出現する単語の出現頻度の時間的変化の類似性を Earth Mover's Distance(EMD) を用いて判定することで，単語間の関連性を分析する手法を提案した [風間 11]．

本稿では，さらに，単語出現頻度のヒストグラムに対してウィンドウを設定し，部分ヒストグラム間の類似性判定とウィンドウの移動を繰り返すスライディングウィンドウ方式により，ツイートに用いられる単語間の関係の時系列変化を分析する方法を提案する．実際に，東日本大震災発生前後の約 3 億ツイートから震災に関係すると思われる単語を抽出し，その出現頻度が上位の単語から「地震」と「原発」の 2 つの単語を選び，その単語と関連語及びその単語間の関係の時系列変化から，現実の事件や議論に Twitter がどのような影響を受けたかを分析する．

2. 東日本大震災と Twitter

2.1 Twitter について

Twitter は，最大 140 文字のつぶやき（ツイート）を，そのフォロワーが閲覧するコミュニケーションサービスである．ツイートは，随時更新されるタイムライン（timeline）上に時系列順に表示され，ほぼリアルタイムで知ることができる．また，SNS の友人関係は双方向で相手の承認が必要だが，Twitter のフォロー関係は一方向で相手の承認が不要なので気軽・容易に構築でき，一方的に知っている著名人とも繋がることができる．さらにリプライ（reply）と呼ぶ特定多数への返信機能や，リツイート（retweet）と呼ぶ情報拡散機能により，情報が急速に広範囲に拡散される特徴を持つ．

連絡先: 風間 一洋，NTT 未来ねっと研究所，〒 180–8585 東京都武蔵野市緑町 3–9–11，kazama@ingrid.org

2.2 震災時の Twitter の利用形態

東日本大震災時の Twitter の主な利用形態は，震災発生時の情報交換手段と震災後の情報拡散手段の 2 種類に分類できる．

震災発生時には，東北・関東地方を中心とした建物の被害や通信回線の切断，大規模な停電，輻輳回避のための通信規制，サーバの過負荷などの理由により，被災地の情報通信インフラに深刻な問題が発生したが，携帯電話やスマートフォンのパケット通信はほとんど規制されなかったために，インターネット上の電子メールやソーシャルメディアが情報交換手段として活用された．さらに，首都圏の交通機関不通により発生した帰宅困難者の情報交換手段としても Twitter が活用された．

震災発生後は，Twitter のフォローの容易さとリツイートによる情報拡散能力に注目して，マスメディアに頼らない情報拡散手段として地方自治体や個人に活用された．ただし，誤った情報やデマも素早く拡散されたことが社会問題になった．

2.3 情報伝播と同期・連鎖現象

実世界やネットワーク上で何らかの大きなイベントが起こった場合には，Twitter 上の複数のユーザから類似した内容が並行にツイートされる現象が観測される．たとえば，地震のような広範囲で同時に体験できるイベントは，膨大な数のユーザによってほぼ同時にツイートされる．また，ユーザがテレビやニュースサイトなどのマスメディアから原発についての新しい情報を知った場合には，その内容に触発されたユーザが非同期にツイートすることでフォロワーに拡散するだけでなく，さらにリプライやリツイートなどの手段によりフォローネットワーク上を連鎖的に情報が伝播していく．

本稿では，このような情報の伝播に伴う同期・連鎖現象に着目する．Twitter 上で，あるトピックがどの程度話し合われているかは，そのトピックを示す単語が単位時間に出現するツイート数から類推できる．その単語出現頻度の時間的変化は，ツイートの拡散・伝播の進行と共に急増し，その後収束するというライフサイクルを持つ．特に東日本大震災のような大規模災害では，地震，津波，停電のような互いに関連する複数のイ

ベントが，並行して繰り返し，または継続的に発生することから，単語出現頻度の時間的変化パターンは単語ごとに固有のパターンとなる．ただし，同じトピックに関する単語は，ツイートの出現頻度の時間的変化が似てくると仮定できるので，その類似性から単語群の関連性を求めることができる．

3. 関連研究

災害やイベント時の Twitter 上の単語出現頻度の時間的変化に関する分析が行われている．総務省の平成 23 年度情報通信白書では，被災地域の自治体とマスメディアが利用している Twitter アカウントのツイート数とフォロワー数の推移や,「コスモ石油」と「コスモ石油　デマ」という 2 種類のクエリで検索した時の検索数の変化から震災関係デマ情報の流布と収束を分析した [総務省 11]．三浦は，東日本大震災における Twitter のリツイート比率と感情表現の変化を分析した [三浦 12]．Sakaki らは，Twitter をソーシャルセンサーと見なして，イベントに関するリアルタイムの相互作用を分析し，地震や台風の発生と位置を推定する手法を提案した [Sakaki 10]．本稿では，Twitter のリプライやリツイートで明示的に示されない経路を含めた情報の伝播を，その情報のトピックに関連する単語群の出現状況から推定するために，既存研究では着目されなかったツイートに出現した単語出現頻度の時間的変化の類似性から同一のトピックを表すと思われる単語群を求める．

4. 単語間の関係の時系列変化の分析

4.1 データセット

3 月 5 日から 24 日の間に，Twitter API[*1]を用いて 200 件以上日本語でツイートしたアクティブなユーザのツイートを収集し，さらに収集漏れを減らすために，後日各ユーザに対して再収集し，それをデータセットとして使用した．200 件は，Twitter API の呼び出し 1 回で取得できる最大ツイート数である．データセットの規模は，ツイート数が 332,414,837 件，ユーザ数が 1,091,741 人である．データセットには，ツイート ID（64 ビット整数），ツイートしたユーザのスクリーン名，本文，ツイート元．ツイート時間，リプライ先のツイート ID，リプライ先のスクリーン名が含まれる．

4.2 震災に関係する名詞の抽出

ツイート本文を Mecab[Kudo 04][*2]で日本語形態素解析して，非自立，数，接尾，ナイ形容詞語幹を除く名詞を抽出した．なお，新しい用語も網羅できるように，標準の IPA 辞書に加えて，はてなキーワード[*3]から作成した辞書を用いた．はてなキーワードは (株) はてなが提供する共有辞書サービスであり，はてなダイアリーで使われたキーワードから自動的にリンクされることから積極的に保守され，流行語や固有名詞，複合語に強い．はてなキーワード辞書の登録単語数は 262,325 語，ツイート本文から抽出した名詞は 463,919 語である．

出現ツイート数の上位には「http://」,「人」のように震災と関連がない名詞も含まれ，さらに出現ツイート数が極端に少ない名詞が 1 回が 30,531 語（6.6%），10 回以下が 110,199 語（23.8%），100 回以下が 210,715 語（45.4%）とかなりの割合を占める．そこで，東日本大震災に深い関連があり，使用頻度がある程度多い名詞だけを対象にするために，次の 3 つの条件を満たす名詞 4,110 語を抽出した．

1. 地震発生から一週間以内の出現ツイート数が 1,000 件以上
2. 一日の出現確率がピークの日が地震発生から一週間以内
3. ピークの日の出現確率が，地震発生前の 10 倍以上

4.3 単語出現頻度の時間的変化による関連性の判定

通常は，単語間の関連性を単語共起から判定することが多いが，Twitter の場合は 140 字制限があるために，どうしても関連があると判定される単語が限定されるという問題がある．

そこで本稿では，単語間の関連性を出現頻度の時間的変化の類似性から判定する [風間 11]．単語出現頻度の時間的変化の類似性は一種の時間的な共起であり，多くのユーザに関連があるイベントが発生した場合に，それに関するツイートとリプライ，リツイートが同期・連鎖することで，顕著な単語出現頻度の変化が生まれる．そこで，逆に単語出現頻度の時間的変化が類似している名詞群が発見できれば，それらは同一のイベントに起因するトピックを表している可能性が高く，単語間の関連性を単語出現頻度の時間的変化から次のように判定する．

4.3.1 単語出現頻度のヒストグラムの作成

最初に与えられたデータセットのツイートを一定の時間幅で分割し，各タイムスロット t に単語が出現するツイート数を調べてヒストグラムを作成する．出現数ではなく出現ツイート数を用いる理由は，ツイート中の単語の繰り返しなどによるバイアスを除去するためである．

4.3.2 Earth Mover's Distance による類似度計算

ヒストグラムの類似度指標として Earth Mover's Distance（EMD）を用いた [Rubner 00]．EMD は 2 つの分布間の距離の指標であり，類似画像検索において画像の色と位置の分布の両方を同時に考慮できるために従来手法に比べて高い性能を示した．本稿では，単語の出現ツイート数と出現時刻という二つの類似性を，同一の距離尺度で同時に判定するために用いた．さらに Twitter のツイートの同期・連鎖に存在する時間差にロバストな点が適している．

EMD では，二つのヒストグラムの間の距離の計算を輸送問題と見なし，ヒストグラムの一方を一定の供給量を持つ複数の供給地，他方を一定の需要量を必要とする需要地として，各供給地から需要地までの輸送コストとしてヒストグラム間のユークリッド距離が与えられた際に，需要地の需要を満たすように供給地から需要地へ最小の輸送コストで荷物を輸送する方法を探し，距離を求める．$EMD(P,Q)$ は供給量ベクトル P，需要量ベクトル Q に対して，次のように定義できる．

$$EMD(P,Q) = \frac{\min_{\{f_{ij}\}} \sum_{ij} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} \tag{1}$$

ただし，以下の制約条件を満たすものとする．

1. 供給地から需要地の 1 方向にしか輸送されない．

$$f_{ij} \geq 0 \tag{2}$$

2. 供給地 i から供給できる容量は供給量 P_i を超えない．

$$\sum_j f_{ij} \leq P_i \tag{3}$$

3. 需要地 j が受け入れる容量は需要量 Q_j 以下である．

$$\sum_i f_{ij} \leq Q_i \tag{4}$$

*1 http://apiwiki.twitter.com/
*2 http://mecab.sourceforge.net/
*3 http://d.hatena.ne.jp/keyword/

4. 供給地から移動できる最大総輸送量は，供給量と需要量の総和の小さい方の量である．

$$\sum_{i,j} f_{ij} = min(\sum_{i} P_i, \sum_{j} Q_j) \quad (5)$$

なお，f_{ij} は供給地 i から需要地 j への輸送量，d_{ij} は供給地 i から需要地 j の間の地上距離（ground distance）である．

本稿では，地上距離に閾値を設定して計算対象となる輸送経路数を削減することで，EMD をさらに高速化した Pele らの $\widehat{EMD}$ アルゴリズムを用いた [Pele 09][*4]．

4.3.3 単語間の関係性の時系列変化の分析

単語の組の関係性の時系列変化は，スライディングウィンドウ方式を用いて分析する．これは，各単語のウィンドウサイズ w に含まれる部分ヒストグラムを $\widehat{EMD}$ で比較してから，そのウィンドウを距離 d だけ移動し，それを対象期間が終了するまで繰り返し，各ウィンドウ期間の単語の組の関連性を示す $\widehat{EMD}$ の値のベクトルを得る手法である．複数の単語の組に適用することで，ある単語の関連語を求めたり，単語集合間の相関関係を求めることができる．

5. 評価

5.1 評価手法

震災に関係する名詞のうち出現ツイート数が上位の名詞は，震災に密接に関係していると考えられるので，出現ツイート数上位の名詞と他を比較し，抽出されたヒストグラムが類似している名詞群を評価する．なお，抽出した名詞の出現ツイート数の上位 10 件は「地震」,「情報」,「停電」,「無事」,「心配」,「被災」,「原発」,「連絡」,「節電」,「避難」だが，本手法で求めた「地震」の関連語が,「情報」,「無事」,「停電」,「心配」,「被災」になるように関連性が強い語が近い順位になることが多いので，なるべく異なるトピックを示す名詞を選択するのが望ましい．そこで,「地震」と「原発」を以後の評価で使用した．

5.2 単位時間のツイート数の変動特性

まず最初に，今回評価に使用する「地震」と「原発」の単位時間の出現ツイート数の変動を図 1 に示す．出現頻度の変動の適度な平滑化を考えて，タイムスロット幅は 10 分にした.「地震」の場合は，3 月 11 日の東日本大震災 (M9.0) 以外にも大きなピークが頻繁に見られるが，これらは 9 日の三陸沖 (M7.2)，12 日の新潟中越 (M6.6)，13 日の茨城沖 (M6.2)，15 日の静岡東部 (M6.0) における地震の発生と合致している．これは，地震は広範囲のユーザ達がほぼ同時に体験することができるイベントなので，ツイートが同期するからだと考えられる．これに対して「原発」の場合には，そのようなピークがほとんど見られない．これは，ユーザが直接原発事故を見ることはできないのでマスコミやソーシャルメディアを経由しなければいけない上に，地震の被害により状態の把握や連絡が困難だったからだと考えられる．このように，単語が示すイベントの性質によって，異なる時間的特性を示すことがわかる．

5.3 関連名詞の時系列変化

本手法で求めた 3 月 11〜23 日の各日の「地震」と「原発」の関連名詞を表 1 と表 2 に示す．ウィンドウサイズと移動量を 24 時間とした理由は，出現頻度の時間的変化の類似性の判定には，ある程度の時間幅が必要なこと，そして人間生活の 24 時間周期の整数倍にすると処理が容易になるからである．

*4 http://www.cs.huji.ac.il/~ofirpele/FastEMD/code/

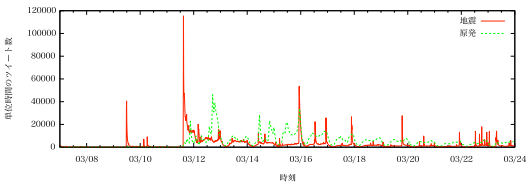

図 1: 名詞の単位時間の出現ツイート数の変動

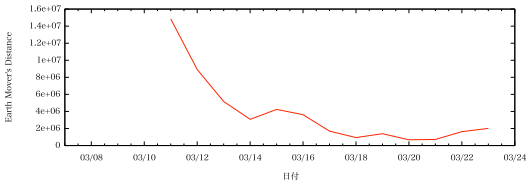

図 2: 名詞間の EMD の変動

全体的には,「地震」の場合は直後は安否を意味する「無事」で，以後は「情報」と常に高い関連があることがわかる．ただし，両方と関連がある名詞も多いが，これは東日本大震災ではさまざまな事柄が同時に多発したために，明確に分離するのが難しいからだと考えられる．

日ごとに見ると,「地震」は 13・14 日が茨城，15 日が静岡のように地震の発生地域「原発」は 11 日の関連語が多彩であるが，これは電源喪失により原子力緊急事態宣言が 19 時頃，水素爆発が起こったのは翌日 15 時半ごろと地震発生とのタイムラグがあり，タイムスロットの区分とうまくマッチしなかったためと考えられる．計画停電に関しては，14 日から開始され，22 日の運用改善のためにグループ細分化 23 日に「放射能」,「放射線」が初めて登場し，順位が下がりつつあった「心配」が急上昇している．これは，同日に内閣府原子力安全委員会が SPEEDI の試算を公開したことで，国民の不安を煽ったと考えられる．

5.4 単語間の関係の時系列変化

次に,「地震」と「原発」の間の関係の時系列変化を分析する．まず，図 2 に名詞間の EMD の変動を示す．図 1 で名詞出現頻度が極端な日は値が大きい．これは，各名詞のイベントの発生時は単独の話題として語られたり，たんに個々につぶやくだけで終わる傾向があるのに対して，それ以外の日は震災関連の話題としてさまざまな要素を考慮して発言されたり，さらに複数人で議論する傾向があるからだと思われる．

さらに，図 3 に各順位までの関連名詞の重複数を示す．各名詞のイベント発生時は関連名詞の重複が少なく，それ以外の日は重複が多いが，これも図 2 と同様な理由からだと考えられる．これから，Twitter における単語の使われ方は，現実世界の事件やリアルタイムな相互作用の影響を受けることがわかる．

6. おわりに

本稿では，Twitter のツイートの同期性・連鎖性に着目し，単語出現頻度のヒストグラムに対してウィンドウを設定し，$\widehat{EMD}$ を用いた部分ヒストグラム間の類似性判定とウィンドウの移動を繰り返すスライディングウィンドウ方式で，震災発生前後の約 3 億ツイートから顕著にツイートされた「地震」,「原発」の日ごとの関連名詞の変化を分析して，現実の事件や議論が

表 1: 「地震」の関連名詞

日付	1	2	3	4	5	6	7	8	9	10
3/11	無事	情報	避難	心配	津波	余震	連絡	拡散希望	確認	拡散
3/12	無事	情報	心配	電気	連絡	節電	余震	拡散	拡散希望	確認
3/13	情報	節電	被災	募金	無事	被災地	電気	茨城	拡散希望	連絡
3/14	情報	節電	被災	計画停電	電気	状況	被災地	茨城	無事	募金
3/15	情報	停電	原発	静岡	無事	被災	心配	被災地	募金	節電
3/16	停電	情報	原発	被災	被災地	福島	状況	心配	無事	募金
3/17	情報	被災地	原発	被災	福島	節電	物資	避難	電気	状況
3/18	情報	被災	原発	被災地	停電	節電	福島	支援	募金	電気
3/19	情報	原発	福島	募金	被災	被災地	震災	停電	避難	節電
3/20	情報	福島	原発	被災	震災	被災地	募金	避難	支援	節電
3/21	情報	福島	原発	被災	被災地	震災	募金	停電	心配	支援
3/22	停電	情報	福島	原発	計画停電	被災	震災	節電	状況	支援
3/23	情報	停電	福島	水	原発	心配	震災	放射能	被災	余震

表 2: 「原発」の関連名詞

日付	1	2	3	4	5	6	7	8	9	10
3/11	協力	一覧	渋滞	運行	帰宅困難者	被災	近辺	全線	交通	被災地
3/12	確認	避難	福島	電気	拡散希望	状況	不安	拡散	被災	節電
3/13	避難	余震	安否	協力	不安	災害	福島	方々	家族	復旧
3/14	確認	報道	無事	募金	東電	心配	連絡	避難	拡散希望	計画
3/15	情報	被災	被災地	無事	心配	状況	募金	節電	東電	福島
3/16	情報	被災地	被災	福島	状況	心配	無事	募金	節電	避難
3/17	被災	被災地	福島	避難	物資	情報	節電	状況	電気	募金
3/18	被災	被災地	節電	福島	情報	支援	募金	電気	状況	震災
3/19	情報	福島	募金	被災	被災地	震災	停電	避難	節電	心配
3/20	情報	福島	被災	震災	被災地	募金	避難	支援	節電	心配
3/21	福島	情報	被災	被災地	震災	募金	停電	支援	心配	ガソリン
3/22	被災	計画停電	震災	福島	情報	節電	状況	支援	被災地	心配
3/23	心配	震災	福島	放射能	被災	状況	被災地	放射線	支援	確認

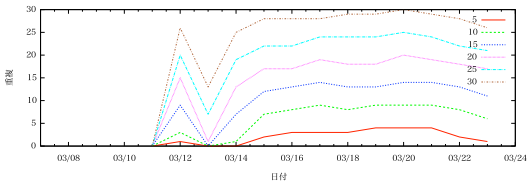

図 3: 各順位の関連名詞の重複

Twitter に与えた影響を分析した.

今後の課題は，TDT (Topic Detection and Tracking) 技術を用いた注目単語とその持続期間の自動検出，データと辞書の整備なども含めた関連単語群の抽出精度の向上，単語のネットワーク構造の時系列変化の分析，定量的評価である．単語出現頻度の変化だけで関連性が求められるので，たとえば注目単語とあらかじめ用意した固有表現，感情表現・顔文字などのあらかじめ用意した辞書の登録語との関連性を分析するのは容易である．さらに，Twitter ネットワークにおける関連単語群の分布の変化を調べることで，Twitter のリプライやリツイートで明示的に示されない経路を含めた情報伝播の推定を試みる．

謝辞

Twitter 検索のために収集したツイートアーカイブを提供して頂いた，クックパッド株式会社の兼山元太氏に感謝します.

参考文献

[風間 11] 風間 一洋, 鳥海 不二夫, 篠田 孝祐, 榊 剛史, 栗原 聡, 野田 五十樹：名詞出現頻度の時間的変化に着目した東日本大震災時の Twitter のトピックの分析, in *WebDB Forum 2011* (2011)

[Kudo 04] Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230–237 (2004)

[三浦 12] 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ーそのときツイッターでは何が起こったかー, 電子情報通信学会誌, Vol. 95, No. 3, pp. 219–223 (2012)

[Pele 09] Pele, O. and Werman, M.: Fast and Robust Earth Mover's Distances, in *Proceedings of 2009 IEEE 12th International Conference on Computer Vision*, pp. 460–467 (2009)

[Rubner 00] Rubner, Y., Tomasi, C., and Guibas, L. J.: The Earth Mover's Distance as a Metric for Image Retrieval, *International Journal of Computer Vision*, Vol. 40, No. 2, pp. 99–121 (2000)

[Sakaki 10] Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors, in *Proceedings of the 19th International Conference on World Wide Web*, pp. 851–860 (2010)

[総務省 11] 総務省：平成 23 年度情報通信白書, `http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h23/pdf/` (2011)