2B3-NFC-02a-3in

ソフトクラスタリングを用いた災害情報の分類

Classification of Information in Disaster by SoftClustering

馬場　正剛 *1　鳥海　不二夫 *1　榊　剛史 *1　篠田　孝祐 *3　栗原　聡 *3
Seigo Baba　Fujio Toriumi　Takeshi Sakaki　Kosuke Shinoda　Satoshi Kurihara

風間　一洋 *4　野田　五十樹 *5　大橋　弘忠 *1
Kazuhiro Kazama　Itsuki Noda　Hirotada Ohashi

*1東京大学 The University of Tokyo
*2電気通信大学 The University of Electro-Communications
*3和歌山大学 Wakayama University
*4産業技術総合研究所 The Nastional Institude of Advanced Industrial Science and Technology

During a disaster, appropriate information must be collected. For example, survivors require information about shelter locations. Rescuers need information about donating money. However, collecting such localized information is difficult from mass media because they generally provide information for the general public. On the other hand, social media can attract more attention than mass media under these circumstances since they can provide such localized information. There are a lot of tweets, so classification of tweets is necessary. Some tweets have more than two topics. For example, a tweet about volunteer is important for victims and rescuers, thus, it is required to classify such kind of information into two classes at the same time. In this paper, we classified tweets posted in disaster. We linked tweets based on retweets to make a retweet network and applied network-soft clustering to the network in order to classify tweets to more than one cluster.

1. はじめに

災害時には個別に必要とする情報の取得が重要である．例えば, 被災者は避難所や被災直後の行動に関する情報によって安全を確保でき, 救援者はボランティアや募金の情報によって, 救援を行える．しかし, このような個別に必要とされる情報は TV, 新聞などのマスメディアからは取得が難しい．一方で, ソーシャルメディアの 1 つである Twitter は, 災害時の個別な情報源として有用であったとの報告が多数存在する [Mendoza 10],[Miyabe 11],[Sakaki 10].

しかしながら, Twitter には多数の投稿 (Tweet) が存在するため, 話題毎に分類されることが必要となる．例えば, 避難所の案内や生活アドバイスは被災者向けの情報として分類され, 支援物資や募金の案内は救援者向けの情報として分類される必要がある．また, Tweet によっては複数の話題に関しており, 択一な分類が適さない場合も存在する．例えば, 炊き出しに関する情報は, 択一的でなく, 被災者向けに関する情報, 救援者向けの情報の両方に分類させる必要がある．

Tweet 分類に関する研究としては, リツイートに注目したネットワーク構造を用いた分類手法である [鳥海 13], [馬場 14] があるが, これらは択一的分類であり, Tweet が複数の話題に関すると分類されない．そこで, 本研究では, [鳥海 13], [馬場 14] の手法に基づいて, リツイートネットワークを構築し, [Zhang 07] のネットワークソフトクラスタリングの手法を用いることで, Tweet が複数のクラスタに所属することを許すクラスタリング, すなわち Tweet のソフトクラスタリングを行う．

連絡先: 馬場正剛，東京大学工学系研究科システム創成学専攻，〒 113-8656 東京都文京区本郷 7-3-1 工学部 8 号館 526，TEL: 03-5841-6991，E-mail:baba@crimson.q.t.u-tokyo.ac.jp

2. Tweet のソフトクラスタリング

2.1 利用データ

使用したデータは，[馬場 14] で扱ったデータと同様である．すなわち, 2011 年 3 月 5 日から同 3 月 24 日までの 19 日間に投稿された日本語の公式リツイートされた Tweet のログデータである．データに含まれる総 Tweet は 30,607,231 件である．なお, ある程度以上の規模で拡散された情報のみを扱うため, 今回はリツイートされた回数が 100 回以上の Tweet のみを対象としてリツイートネットワークを構築した．100 回以上リツイートされた Tweet は 34,860 件であった．

2.2 リツイートネットワークの構築

鳥海らの手法 [鳥海 13] に基づき, 二部グラフを使用して, リツイートネットワークを構築した．ある 2 つの Tweet に対して同時にリツイートを行ったユーザが複数人存在した場合, 彼らはその 2 つの Tweet に類似した興味を持っていたと考えられ, それらの Tweet には内容の類似性があると推定される．このとき, リツイートしたユーザの重複度が高い Tweet 同士をリンクで結ぶことで, 内容の類似性に基づいたリツイートネットワークの構築が可能である．このリンクの接続手法は Small の共起の手法 [Small 73] に基づいている．

2 つの Tweet t_i,t_j をリツイートしたユーザ群 U_i,U_j のユーザ群重複率は, Jaccard 係数 [Frakes 92] を用いて次のように求められる．

$$O_{ij} = \frac{|U_i| \cap |U_j|}{|U_i| \cup |U_j|} \quad (1)$$

ユーザ群重複率 $Q_{i,j}$ が閾値 $th = 0.05$ 以上の 2 つのリツイートをリンクで結ぶことで, リツイートネットワークを構築した．また, 他の Tweet とリンクで結ばれてない Tweet, すなわち独立したノードは, 今回は分析の対象から除外した．リツイートされた回数が 100 以上の Tweet の内, 重複率 O_{ij} が

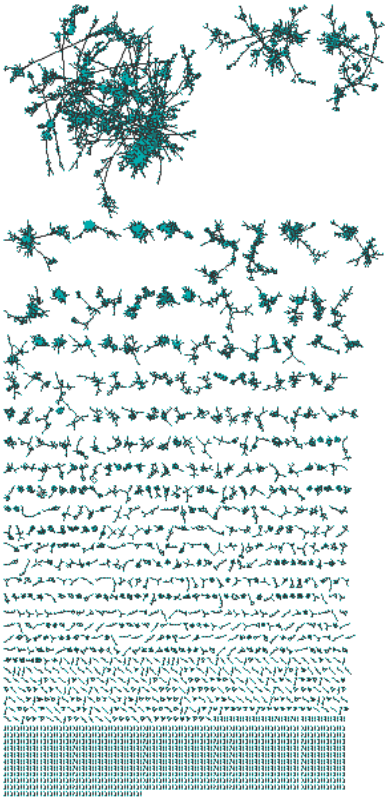

図 1: リツイートネットワーク

$th = 0.05$ 以上のペアを持つ Tweet は 11,494 件であり, リンク数は 30,363 本であった.

ここで, 得られたネットワークを図 1 に示す. ノードは Tweet を示し, エッジはリツイートしたユーザの重複率が高い Tweet 同士であることを示している. コンポーネントに含まれるノード数は様々であり, 下部にはノード数が少ないコンポーネントが存在する一方で, ノード数が非常に多いコンポーネントが上部に存在している.

また, ノード数が上位 10 件のコンポーネントを表 1 にまとめる. 図 1 での左上部のコンポーネントが最も多くノードを含んでおり, その数は 2234 ノードである.

表 1: ノード数が上位 10 件のコンポーネント

順位	ノード数
1	2234
2	347
3	288
4	159
5	142
6	116
7	111
8	109
9	104
10	98

2.3 ネットワークソフトクラスタリング

図 1 のコンポーネントの中には, 多数のノードが所属しているコンポーネントも多く存在する. 例えば, 左上部のコンポーネントである. このようなコンポーネントには様々な話題の Tweet が混在しており, 複数の話題に関する Tweet も存在していると考えられる.

そこで, コンポーネントに所属している多数の Tweet を更に分類すべく, ネットワークソフトクラスタリングを行う. 用いたソフトクラスタリング手法は Zhang[Zhang 07] の手法に基づいている.

Zhang の提案したソフトクラスタリングのアルゴリズムは次のようである.

- クラスタ数の上限を K, ネットワークの隣接行列を $(a_{ij})_{n\times n}$, クラスタへの所属閾値を λ とする.

1. Spectral Mapping
   (a) 対角行列 $D = (d_{ii}), d_{ii} = \sum_k a_{ik}$ を計算
   (b) 一般化固有値問題 $Ax = tDx$ を解き, 上位 K 個の固有ベクトルから固有ベクトル行列 $E_K = [e_1, e_2, , , , e_K]$ を生成
2. Fuzzy c-means
   (a) クラスタ数 k$(2 \leq k \leq K)$ を選択
   (b) E_K から $E_k = [e_2, e_3, , , , e_k]$ を生成
   (c) ユークリッドノルムを用いて, E_k の行ベクトルを単位長に正規化
   (d) fuzzy c-means により E_k の行ベクトルのクラスタリングを行い, 所属行列 U_k を計算
3. 拡張 $\widetilde{Q}(U_k)$ が最大値をとる k と所属行列を決定

Zhang が提案した拡張 $\widetilde{Q}(U_k)$ は $n \times k$ の所属行列 $U_k = [u_1, u_2, , , , u_k]\{0 \leq u_{ic} \leq 1, \sum_{c=1}^{k} u_{ic} = 1, (\text{ただし}, c = 1, ...k, i = 1, ...n)\}$ を用いて次のように表される.

$$\widetilde{Q}(U_k) = \sum_{c=1}^{k} [\frac{A(\overline{V_c}, \overline{V_c})}{A(V,V)} - \left(\frac{A(\overline{V_c}, V)}{A(V,V)}\right)^2], \qquad (2)$$

ただし,
$A(\overline{V_c}, \overline{V_c}) = \sum_{i\in\overline{V_c}, j\in\overline{V_c}} \frac{(u_{ic}+u_{jc})}{2} a(i,j)$,
$A(\overline{V_c}, V) = A(\overline{V_c}, \overline{V_c}) + \sum_{i\in\overline{V_c}, j\in V/\overline{V_c}} \frac{(u_{ic}+(1-u_{jc}))}{2} a(i,j)$,
$A(V,V) = \sum_{i\in V, j\in V} a(i,j)$
$\overline{V_c} = \{i | u_{ic} > \lambda\ , i \in \overline{V}\}$.

この拡張 $\widetilde{Q}(U_k)$ は Newman の Q[Newman 04] の一般化であり, ソフトクラスタリングを適用した結果の良さを表す指標となる.

表 1 での最大コンポーネントである, 2234 件のノードで構成されたコンポーネントへのソフトクラスタリングの適用を行った. 最大クラスタ数 K を決めるために, [馬場 14] で提案された拡張 Newman 法を適用したところ, $k = 31, Q = 0.854$ を得た. ハードクラスタリングである拡張 Newman 法で最適クラスタ数が 31 であるならば, ソフトクラスタリングにおいての最適クラスタ数も高々100 であると考えられるため, 最大クラスタ数を K=100 とした. 所属閾値 λ は, [Zhang 07] での適用例で最も大規模なネットワークを扱った際の λ =0.10 に準拠して, 本稿でも λ =0.10 とした. また, Fuzzy c-means の m は m=2 とし, 初期中心の選定は [Zhang 07] に従い, できるだけ互いに直交するように選定した.

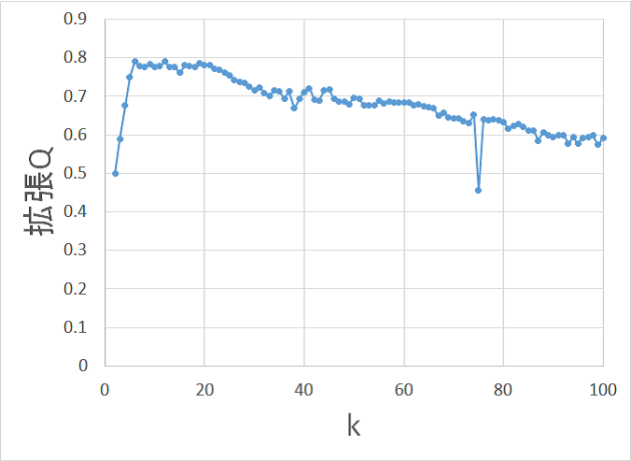

図 2: 拡張 Q とクラスタ数 k の関係

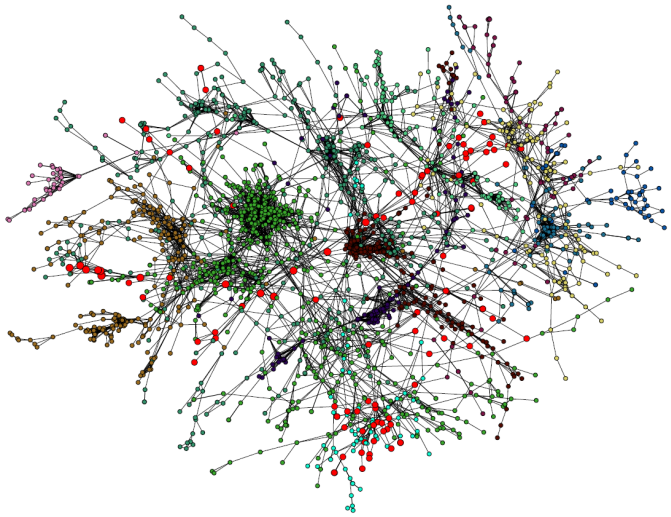

図 3: ソフトクラスタリングされたネットワーク

表 2: クラスタに含まれる情報

クラスタ番号	ノード数	主な発言者	主な内容
0	686	メディア各種	被災者向けの情報
1	134	岩手県庁	岩手県ローカル情報
2	417	NHK 各種	計画停電, 支援物資, 避難所での生活
3	37	有名バンド	震災直後の対応
4	75	NHKNews	news 全般
5	67	ジャーナリスト	放射能, 原発
6	101	首相官邸	国民への呼びかけ
7	128	NHK 生活, 地震速報	ライフライン, 支援物資, 安否確認
8	122	消防庁, NHK 各種部署	被害状況取りまとめ, 震度
9	109	有名女性歌手	震災直後に特化した情報（避難所・避難方法）
10	255	有名バンド, 有名女性歌手	避難所・生活アドバイス
11	228	東大物理学者, 東大病院放射線治療チーム	原発・放射能

表 3: 複数クラスタに所属する Tweet 数

所属クラスタ数	Tweet 数
2	81
3	13

3. 結果

3.1 クラスタリング結果

拡張 Q の計算結果を図 2 に示す. k=12 に拡張 Q は最大値 0.79028 を取ったため, k=12 を最適なクラスタ数として採用した. k=12 における, ソフトクラスタリングの結果のネットワークを図 3 に示す. 色はノードが所属しているクラスタを表している. また, 複数のクラスタに所属しているノードはややノードサイズを大きくし, 赤色にした. 左中部の緑色のノード群が最大クラスタである.

3.2 クラスタの話題

得られた分類結果において, 話題毎にクラスタがわかれていることが分かった. 例えば, 被災者向けの情報（給水所情報, 避難所での生活情報）, 岩手県ローカル情報（岩手県内安否確認情報, 県内交通情報）などとしてまとめられていた. 全クラスタの主な発言者と主な内容を表 2 に示す.

3.3 複数クラスタに所属する Tweet

複数クラスタに所属する Tweet 数を表 3 に示す. 複数クラスタに所属する Tweet の大部分は 2 クラスタに所属している. また, 複数クラスタに所属する Tweet 例を表 4 に示す. 表 4 の 1 つ目の Tweet は被災者向けの生活アドバイス情報であり, クラスタ 0,10 の複数クラスタに所属している. この Tweet は被災者向けの情報であり, 生活アドバイスであるにも関わらず, 択一的分類であるハードクラスタリングにおいては, この Tweet は 1 クラスタのみへの所属が強いられる. しかし, 本手法によっては, 複数のクラスタに所属しているとされ, 被災者向けの情報だけでなく, 生活アドバイスに関する情報であることが明らかになった. また, 表 4 の 2 つ目の Tweet は原発に関する情報であり, 原発周辺で生活している方向けのアドバイスでもある. この Tweet も, 1 つのクラスタに択一的には所属せず, 話題に応じて複数クラスタに所属しており, 原発・放射能に関する情報の中でも, 原発周辺住民向けの情報であることが明らかになった. 同様に, 他の Tweet も話題に応じて, 複数クラスタに所属しており, 情報の特徴が明確になった.

これらより, 本手法では, 話題毎に Tweet を分類するとともに, 複数の話題に関する Tweet は複数クラスタに所属する Tweet として検出されると言える.

表 4: 複数クラスタに所属する Tweet 例

所属クラスタ番号	Tweet
0,10	【被災者の方へ】被災者のために全国の自治体が用意している住宅（19日現在）の一覧リンクです. http://t.asahi.com/1p58　入居期間は3カ月～1年程度. ほとんどが無料です. #jishin
10,11	【福島原発周辺にお住まいの方へ】健康相談ホットライン（0120-755-199）を開設しました.　具体的な除染方法等は, 090-5582-3521 090-4836-9386 080-2078-3308［続く］ #mext #jishin
5,7	立て続けの緊急地震速報でした. これから深夜にかけて余震の際にはいちだんと気をつけて下さい. まずは落ち着いて行動することが大事です. お年寄りの方が近くにいる人はどうぞ助け合って行動して下さい. #nhk #kaigo #jishin
0,4,8	全力拡散. RT @Yoshiteru_Iio: @sasakitoshinao 先ほど, 日本ユニバーサルデザイン研究機構に行ってきました. 物資の集積所の画像をアップしましたので, ご参考までに. http://twitpic.com/4a7xx5
0,4,8	被災地で必要なアイテムはこのサイトの下の方に一覧表になっています. 確認して是非持ち込みを. ／【ユニバ地震対策本部】被災地への救援物資を送付したい方へ http://t.co/t519KmC

4. 結論

本研究では, 震災期間中に投稿された Tweet から, リツイートネットワークを構築し, ネットワークソフトクラスタリングを用いることで, Tweet を分類した. 得られたクラスタは, 話題毎に Tweet が分類されていることを確認し, 話題が択一的に決まらない Tweet は複数のクラスタに所属する Tweet として検出した.

今後の課題としては, 重なり構造だけでなく, 階層構造を検出するクラスタリング手法の適用, 震災時以外のデータでの本手法の妥当性の検証, 計算速度の向上などが挙げられる.

5. 謝辞

本研究で利用した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に感謝する. 本研究の一部は, 日本学術振興会課題設定による先導的人文・社会科学研究推進事業による.

参考文献

[Mendoza 10] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: can we trust what we RT? In Proceedings of the First Workshop on Social Media Analytics -SOMA'10, pages 71-79. ACM Press, July 2010.

[Miyabe 11] M. Miyabe, E. Aramaki, and A. Miura. Use trend analysis of twitter after the great east japan earthquake. In Proceedings of SIG-DPS/GN 2011-DPS-148/2011-GN-81/2011-EIP-53, 2011.

[Sakaki 10] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, WWW'10, pages 851-860. ACM, 2010.

[鳥海 13] Fujio Toriumi, Takeshi Sakaki, Kosuke Shinoda, Kazuhiro Kazama, Satoshi Kurihara, and Itsuki Noda. Information Sharing on Twitter During the 2011 Catastrophic Earthquake. 2nd International Workshop on Social Web for Disaster Management (swdm2013) WWW 2013 Companion Publication pp.1025-1028

[Frakes 92] W. B. Frakes and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Prentice Hall PTR, 1992.

[Newman 04] Clauset, A., Newman, M. E., and Moore, C. Finding community structure in very large networks, *Physical review E*, Vol. 70, No. 6, p. 066111 (2004)

[馬場 14] 馬場正剛, 鳥海不二夫, 篠田孝祐, 榊剛史, 栗原聡, 風間一洋, 野田五十樹, 大橋弘忠: 災害情報の分類の妥当性の評価 (2014)

[Small 73] Small HENRY. Co- citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science, 24(4), 265-269, 1973.

[Zhang 07] Shihua Zhang, Rui-Seng Wang, Xiang-Sun Zhang. Identification of Overlapping Community Structure in Complex Networks using Fuzzy C-means Clustering. PHYSICA A, 374, pages 483-490. 2007