2B4-NFC-02a-3

ネットワーク構造に基づく災害情報の分類

Classification of Information in Disaster Situation from Network Structures

鳥海不二夫[*1] Fujio Toriumi　榊 剛史[*1] Takeshi Sakaki　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]慶應義塾大学/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 The Nastional Institude of Advanced Industrial Science and Technology

Wide-scale disasters such as earthquakes, hurricanes and so on, occur unpredictably. During a disaster, it's important to collect information appropriately to save own lives. However, it is difficult to collect information from mass media, such as TV, Newspapers, which contains information which is of use for the general public. Under the disaster situation, victims require information which shows place of shelters or danger points. Also, not only victims but also rescuers require information of victim location or that of shorted supplies. In this paper, we classify information that was diffused widely on Twitter to clarify what kind of information is required by victims. A bipartite graph that consisted of tweets and retweeted users to classify the retweeted tweets is used for the analysis. By analyzing the network of similar tweets using clusters which calculated by the Newman method, we classified each tweet from the viewpoint of users who have similar interests.

1. 緒言

地震や台風と行った大規模な災害はいつ起きるか分からない．この 10 年に限定しても，スマトラ沖地震 (2004)，ハリケーンカトリーナ (2005)，四川大地震 (2008)，チリ地震 (2010)，東日本大震災 (2011) など数多くの大災害が人々を襲っている．また，マグニチュード 7.0 を超える地震だけでも 2010 年には 24 回，2011 年には 20 回観測されている[*1]．

このような災害時には，情報を正確に素早く集めることが人命を守るために重要となる．しかしながら，新聞やテレビといったマスメディアは一般的な情報を提供することを目的としている．そのため，避難所の場所や被災地に必要な物資など，被災者や救助者が必要としている情報を必ずしも提供していない．

このような状況下で，ソーシャルメディアによる情報の共有が注目されている．特に，ソーシャルメディアの一つであるツイッターによる災害時の情報共有については，多くの報告が存在する [Vieweg 10] [Heverin 10]．

本研究でもツイッターに着目し，東日本大震災時にツイッター上で共有された情報について，ユーザのリツイート行動に基づく情報の分類法を提案する．得られたクラスタを分析し，提案手法によって分類された情報群の特徴を明らかにするとともに，どのような情報がツイッター上に共有されたかを明らかにする．

2. 関連研究

ツイッター上のクラスタリング手法に関する研究としては，Herdağdelen らによるユーザのクラスタリング [Herdağdelen 12] や，リンク構造を用いた重複クラスタリング手法によるユーザのクラスタリング [金川 11]，記事嗜好に基づくユーザのクラスタリング [裕也 10] など，ユーザの分類を目的とした研究が数多くなされている．

また，ツイートそのものの分類手法としては，TweetMotif を用いたトピックごとの分類手法 [Brendan 10]，ハッシュタグを用いた分類手法 [Rosa 11] などがある．一方で，本手法ではツイートを対象として，言語処理やハッシュタグなど情報の内容を利用せずに，リツイートに基づくネットワークのみからクラスタリングを行うことで，必要とするユーザの同一性から情報を分類することを目指している．

3. ネットワーク構造を用いたリツイートの分類

3.1 二部グラフによるネットワークの構築

ある二つのツイートを同時にリツイートしたユーザが複数人いた場合，二つのツイートは共通した興味を持たれる内容を有していると考えられる．そこで，リツイートしたユーザの重複度からツイートはクラスタリング可能であり，それぞれのクラスタの内容から，震災時に共有された情報の種類を明らかにできると考えられる．このように，リツイート関係のみを利用してクラスタリングを行うことで，言語的なクラスタリングでは得られない「必要とするユーザの類似性」によってツイートを分類することが可能である．これによって,「情報の内容」ではない視点から伝播した情報を分析する．

本研究では二部グラフ [Wasserman 94] によって作られたネットワークを用いてツイートのクラスタリングを行う．まず，類似した内容を獲得するために，RT を行ったユーザによる分類を行った．二つのリツイート rt_i, rt_j を取り出し，それぞれのリツイートを行ったユーザ群 U_i, U_j の重複率が高いリツイート同士を隣接リツイートととらえ，ネットワークを構築することで類似リツイートを取り出す．すなわち，ツイートをノード及び重複率の高いツイート同士をつなぐリンクによってネットワークが構築される．なお，ある程度以上の規模で広

連絡先: 鳥海不二夫，東京大学大学院工学系研究科システム創成学専攻，東京都文京区本郷 7-3-1，tori@sys.t.u-tokyo.ac.jp

*1 http://on.doi.gov/7cqeex

まった情報を対象とするため，今回はリツイート数が 100 以上のツイートのみを対象としてネットワークを構築した．

このとき，ユーザ群の重複率 O_{ij} は Jaccard 係数を用いて以下のように求める．

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

重複率 O_{ij} が閾値 th 以上のペアをリンクでつなぎ，ネットワークを構築する．ここで，th の取り方によってネットワークの構造が変化する．が，今回は $th = 0.05$ とした．また，独立したノード，すなわち他のノードと接続していないノードは分析の対象から外した．

さらに，得られたネットワークについて，コミュニティ抽出を行い関係性の深いツイートの集合を獲得する．コミュニティ抽出には，Modurality を基準とする Newman 法 [Clauset 04] を用いた．

3.2 ツイートネットワーク

得られたツイートネットワークを図 1 に示す．

これらのノードをコミュニティ抽出により分割した結果，得られたクラスタについてみる．得られたクラスタ数は 2048 であり，最も大きいコミュニティで 613 ノードが所属している．

各クラスタに含まれるノード数の分布を図 2 に示す．これより，各クラスタに含まれるノード数は概ねベキ分布を形成していることが分かった．

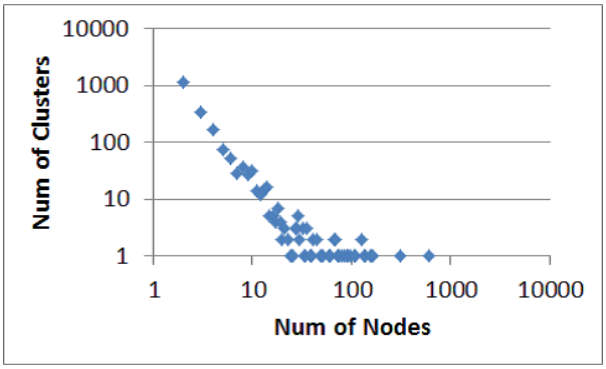

図 2: クラスタのノード数分布

4. ハブアカウントの影響

本研究で提案する情報の分類手法は，多くのユーザが同時にリツイートしたツイートは，関係のある情報であるという仮定に基づいている．一方で，フォロアの多いハブユーザが存在すると，当該ユーザのツイートは多くのフォロアによってリツイートされるため，クラスタを形成しやすい．また，ハブユーザがリツイートしたツイートもそのフォロアを介して広まりやすい．

そこで，各クラスタを発信者の集中度と，メンション[*2]されたユーザの同一性によって評価する．

4.1 発信者の集中度

発信者の集中度は情報量エントロピーによって分析する．情報量エントロピーは，

$$H(X) = -\sum_i P_i \log P_i \quad (2)$$

によって求められる量である．ただし，P_i はあるクラスタに所属するツイートのうちユーザ i が投稿したものの割合とする．発信者エントロピーが高いクラスタは，様々な人物のツイートによって構成されているクラスタであり，発信者エントロピーの低いクラスタは，特定の人物のツイートによって構成されているクラスタである．

各クラスタについて発信者の情報量エントロピーを求めた結果を図 3 に示す．横軸がエントロピー，縦軸が累積頻度である．これより，全クラスタの 62.4% がエントロピー 0，すなわち一人のユーザによって書かれたツイートがまとめられたものであったことが分かった．同様に，エントロピー 1 未満のツイートが全体の 80%以上を占めていた．

これより，本手法によって分類された多くのクラスタが，単一または少数のユーザによるツイートによって構成されていることが明らかとなった．

ここで，エントロピーが特に小さいクラスタ ($H(X) < 0.5$) の中でノード数の多いものから上位 10 クラスタについて，最も多く情報発信を行っていたユーザを抽出した．その結果を表 1 に示す．その結果，いずれのアカウントもマスメディアの公式アカウントであったり，芸能人のアカウントであったりといわゆる有名人のアカウントであった．これより，特定の情報発信者のツイートがクラスタを形成していることが明らかとなった．

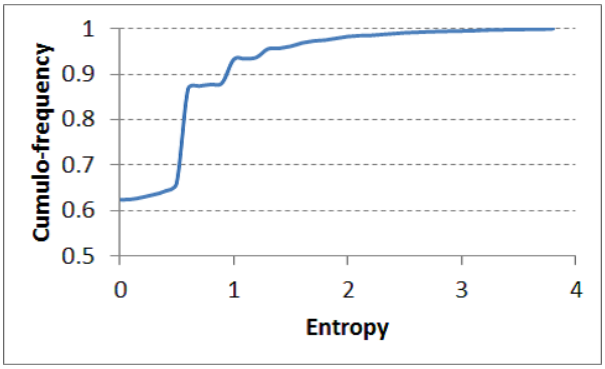

図 3: クラスタのノード数分布

表 1: 最頻情報発信者

Nodes	$H(X)$	Top User Rate	Top User
153	0.434	0.843	東大病院放射線治療チーム
128	0.273	0.953	特撮ヒーロー情報
107	0.146	0.972	歌手
74	0.072	0.986	茨城新聞社
73	0.00	1.00	福岡市長
69	0.409	0.899	NHK 広報
62	0.00	1.00	通信会社社長
50	0.00	1.00	NHK ニュース
45	0.00	1.00	キャラクターボット
41	0.00	1.00	フリーアナウンサー

*2 *@username*

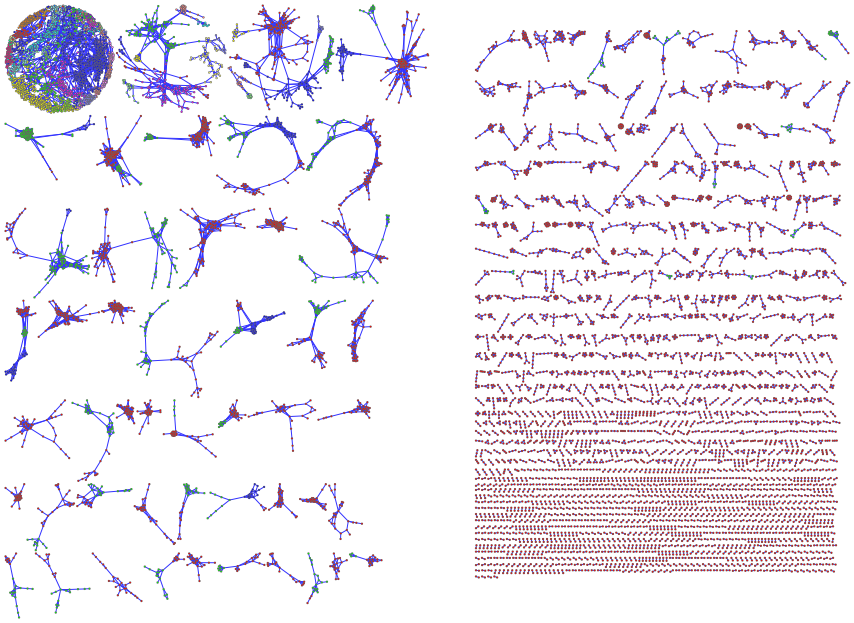

図 1: Retweet Network

4.2 メンション対象の同一性

得られたクラスタの中には，同一人物へのメンションが多く含まれるクラスタが存在している．これは，芸能人のようなハブユーザに向けたツイートを行うことで，より多くの人と情報を共有しようとした結果であると考えられる．一度ハブユーザがリツイートを行えば，当該ツイートはハブユーザがツイートを行ったのと同程度の情報拡散が期待できる．

このようなツイート群を抽出するために，最大メンション占有率 $M(X)$ を用いる．最大メンション占有率は，最も多くメンションされたユーザに対してメンションされたツイート数が，全ツイートに対して占める割合である．

その結果，全体の 22.5% となる 461 のクラスタで，クラスタ内のツイートの 50% で同一のユーザに向けたメンションが行われていたことが分かった．

ここで，占有率の高いクラスタ ($M(X) \geq 0.5$) の中でノード数の多いものから上位 10 クラスタについて，最も多く情報発信を行っていたユーザを抽出した．その結果を表 2 に示す．その結果，いずれのアカウントもマスメディアの公式アカウントであったり，芸能人のアカウントであったりといわゆる有名人のアカウントである．

以上，特定のアカウントの影響力が強いクラスタについて分析を行った．その結果，有名人のツイートおよび有名人がリツイートしたツイート群がクラスタを形成しやすいことが明らかとなった．

実際，得られたクラスタの内 80.0% が $H(X) < 0.5$ または $M(X) \geq 0.5$ を満たしている．したがって，本手法における情報の分類には，ハブアカウントが大きく寄与していることが分かった．

一方で，28 クラスタでハブユーザであったユーザも存在していることから，ハブユーザの影響が強いクラスタであっても，同一クラスタにまとめられるわけではないことが分かる．すなわち，単にハブとなるユーザだけが基準にクラスタが作られているわけではなく，ハブユーザのツイートがさらにユーザの興味に従って分類されていると言える．

表 2: 最頻メンション対象者

Nodes	Mention Rate	Target
107	0.636	歌手
28	0.893	お笑い芸人
27	0.926	お笑い芸人
19	0.526	元 IT 系社長
18	0.667	韓国人歌手
17	0.824	モデル
14	0.643	歌手
13	1.000	脳科学者
13	1.000	元県知事
12	0.667	モデル

5. クラスタに含まれる情報

最後に，得られたクラスタにどのような情報が含まれるかを確認した．全クラスタを網羅するには紙面が足りないため，含まれるノード数が上位 10 クラスタについて含まれる主な情報を表 3 にまとめた．

得られたクラスタごとに話題は限定されており，クラスタリングはおおむね成功していた．全クラスタを確認したところ，その内容は，避難者の受け入れ情報や避難生活者へのアドバイス，計画停電情報，放射線に関する情報などクラスタごとに大まかに分かれている．

一方で，芸能人によるツイートを中心としたクラスタ (ノード数 3，7 位のクラスタ) も存在し，それらのツイートに含まれる情報には統一性はないが，避難物資の不足を訴えるツイートや個人からの情報提供が多かった．

その他，下位のクラスタまで確認した結果，

表 3: 上位クラスタに含まれる情報

Rank	Nodes	Retweeted Count	Information	Common informer
1	613	164216	各種マスメディアによる被災者への情報提供	朝日新聞社会部
2	313	87216	NHK アカウントによる計画停電などインフラに関する情報	ＮＨＫ生活情報部
3	163	55203	女性歌手のツイートおよび当該ユーザのリツイート	有名女性歌手
4	158	41435	岩手県を中心とした東北地方の被災状況	IBC 岩手放, 岩手県広聴広報課
5	153	88118	放射線に関する情報	東大病院放射線治療チーム
6	137	39551	官邸，マスメディアからの被災地へのお知らせ	首相官邸（災害情報）
7	132	37627	バンドボーカルのツイートおよび当該ユーザのリツイート	有名バンドボーカル
8	128	67964	NHK アカウントによる震災直後の情報提供	ＮＨＫ生活情報部
9	128	34760	特撮ヒーローから子供達への励ましのメッセージ	特撮ヒーロー
10	109	25939	自衛隊，政府関係者の対応に関する情報	政治家，一般ユーザ

- 被災者向けのアドバイス
- 地域情報
- 放射線に関する情報
- 原発事故に関する情報
- 計画停電，節電に関する情報
- 政府，東京電力批判
- 震災時の経験談
- 自衛隊，海外からの支援に関する情報

などがクラスタごとの話題として登場していた．

また，被災したペットに関する情報が一つのクラスタにまとめらている，原発問題と放射能問題が異なるクラスタに分けられている，など言語処理では困難な情報の分類が実現されていることが確認された．

6. 結言

本研究では，二部グラフとコミュニティ分類の手法を用いて，震災時に Twitter 上に投稿された情報の分類を試みた．提案手法によってツイートを話題ごとに分類することに成功した．それらのクラスタは主に Follower の多いユーザがハブとなり情報がまとめられていることを確認した．ハブとなるユーザは，芸能人，マスメディアのアカウントなどである．一方で，各クラスタは単にハブユーザがツイートした情報というだけでまとめられているわけではなく，同一アカウントのツイートでも，内容によって異なるクラスタに分類されていることが確認された．さらに，具体的にクラスタ内に存在するツイートを確認し，適切な分類が行われていることを確認した．

本手法を応用することで，震災時などに Twitter などソーシャルメディアからの情報をより素早く分類することが可能になると期待される．一方で，本論文ではクラスタの分類結果の妥当性の評価を行えていない．これらの分類結果の妥当性を評価することは今後の課題となる．

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に感謝する．また，本研究は科研費 (24300064) の助成を受けて行われたものである．

参考文献

[Brendan 10] Brendan O ’ Connor, Krieger, M., Ahn, D. : Tweetmotif: Exploratory search and topic summarization for twitter, *Proceedings of ICWSM*, pp. 2–3 (2010)

[Clauset 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)

[Herdağdelen 12] Herdağdelen, A., Zuo, W., Gard-Murray, A., and Bar-Yam, Y.: An Exploration of Social Identity: The Geography and Politics of News-Sharing Communities in Twitter, *arXiv preprint arXiv:1202.4393* (2012)

[Heverin 10] Heverin, T. and Zach, L.: Microblogging for Crisis Communication: Examination of Twitter Use in Response to a 2009 Violent Crisis in Seattle-Tacoma, Washington Area, in *Proceedings of the 7th International ISCRAM Conference* (2010)

[Rosa 11] Rosa, K. D., Shah, R., Lin, B., Gershman, A., and Frederking, R.: Topical clustering of tweets, in *Proceedings of SIGIR Workshop on Social Web Search and Mining* (2011)

[Vieweg 10] Vieweg, S.: Microblogged Contributions to the Emergency Arena: Discovery, Interpretation and Implications, in *Computer Supported Collaborative Work* (2010)

[Wasserman 94] Wasserman, S. and Faust, K.: Social Network Analysis: Methods and Applications, in *Structural Analysis in the Social Sciences*, Vol. 8, pp. 299–302, Cambridge University Press (1994)

[金川 11] 金川 元信, 大豆生田利章 : ソーシャルネットワークにおけるリンク構造を用いた重複クラスタリング手法の提案, *DEIM Forum 2011* (2011)

[裕也 10] 裕也 眞野, 俊弘 青山 : ミニブログユーザの記事嗜好を用いたクラスタ発見, 日本高専学会誌 : journal of the Japan Association for College of Technology, Vol. 15, No. 3, pp. 43–46 (2010)