Twitter ハッシュタグの自律的な組織化について

川端 智久† 白井 靖人‡ 石川 博‡

†静岡大学大学院情報学研究科 〒432-8011 静岡県浜松市中区城北 3-5-1
‡静岡大学情報学部情報科学科 〒432-8011 静岡県浜松市中区城北 3-5-1
E-mail: †gs10013@s.inf.shizuoka.ac.jp, ‡{shirai,ishikawa}@inf.shizuoka.ac.jp

あらまし フォークソノミーにおける共通の問題は，同義語やタグの抽象度の異なりによって適切なタグが一つに定まらないことである．Twitter の ハッシュタグにもこの問題は当てはまるが，ユーザによる継続的な利用を通して適切にタグが使い分けらていく事例が見られる．このようなハッシュタグの組織化は，同義タグが一つに集約される「収束」と，一つのハッシュタグ内で混在する話題に個別のハッシュタグが与えられる「分化」の二つのプロセスで成り立つと考えられる．震災関連ハッシュタグを対象に，組織化にどのような傾向があるかを分析した．分化について，分化の直後はハッシュタグの共起率が高いため新しいハッシュタグを周知する効果があり，次第にあるいは急激に共起率が減少して目的に特化して使われるようになる傾向がある．とくに共起率の減少とツイート数の増加が同時に起こったとき，ユーザの関心が大きく変化したことを読み取れる．

キーワード Twitter，ハッシュタグ，タグ組織化，時系列分析

1. はじめに

マイクロブログ Twitter ではハッシュタグとよばれるタグづけ機能が利用されている．ハッシュタグはキーワードの先頭に「#」(hash mark)をつけて記述される．ハッシュタグには検索結果へのリンクが張られ，そのハッシュタグを含むツイートを一覧できる．ハッシュタグは多数のユーザが自由にタグづけをおこなうフォークソノミーの一種といえるが，ツイート投稿者本人が投稿時にしか記述できないことや，テレビ番組実況のようにリアルタイム性の高い用途でも使われることが特徴である．

フォークソノミーの利用や分析における共通の問題は，同義語や多義語，またタグの抽象度の違いによって適切なタグが一つに定まらないことである．表記揺れによって同じ意味のタグが複数並立して使われることや，いくつかの異なる話題に一つのタグが混同して使われることが問題になる．

ハッシュタグについてもこの問題は当てはまる．しかし，ユーザによる継続的な利用を通して適切にタグが使い分けらていく事例も見られる[1]．たとえば東日本大震災の直後では，様々な話題が「#jishin」という単一のハッシュタグでツイートされていたが，次第に原発や地域ごとの情報に関するハッシュタグ，復興支援を目的としたハッシュタグなど，話題や目的に応じたハッシュタグが増えていった．震災関連ハッシュタグは Twitter が公式に推奨ハッシュタグを告知したものもあるが*，ほかにもさまざまなハッシュタグがユーザ主体で提唱され，使い分けが進んだ．

本稿ではこれをハッシュタグの自律的な組織化とよび，このような組織化にはどのようなパターンがあり，またそれがどのような要因で起こるかを分析する．タグの並立や混同という問題にユーザがどう対処するかを明らかにすることで，タグの新しい分析手法や組織化を促す方法の発見に寄与できると考える．

2. 関連研究

村井[1]は，東日本大震災に関するハッシュタグを分析し，ハッシュタグが自律的に組織化することを発見した．ハッシュタグごとの話題変化やハッシュタグ同士の関係性について，頻出名詞抽出やネットワーク分析を用いて分析している．本研究では，このような組織化がどのようなパターンや傾向をもつのかを特定し，組織化の生じる要因を明らかにすることを目指す．

本研究ではユーザが時間をかけて自主的にタグを使い分けていく自律的な組織化を分析対象とするが，それとは対照的に，大量に収集したタグづけデータを分析することでタグの集約や階層化をおこなう機械的な組織化も研究されている．

多田ら[4]は，タグを使用したユーザ，タグが使用されたリソース，タグと同時に使用された共起タグの三つの観点からタグ間の類似度を算出し，複数の同義タグを一つのタグに集約する手法を提案している．丹羽ら[5]は，ユーザベース共起率とドキュメントベース共起率の組み合わせたクラスタリングによって同義タグを判定する手法を提案している．これは，一つのドキュメントに複数の同義タグがつくことはあっても，一人のユーザが複数の同義タグを使い分けることはないだろう，という仮説に基づく．また同義タグだけでなく，「windows」と「mac」のように同じレベルで競合関係にある Conflict 関係，意味的な相関度が高い Relevant 関係も判定する．

タグの階層化について，村上ら[6]は動画投稿サイト

* Twitter ブログ: 東北地方太平洋沖地震に関して
http://blog.jp.twitter.com/2011/03/blog-post_12.html

を対象に，谷田川ら[3]はハッシュタグを対象に試みている．これらの研究では階層化の手法に Li ら[2]の提案した ISR 手法を用いる．杉本ら[7]も同様のアプローチでタグの階層関係抽出を試みている．

以上のようなタグの機械的な組織化に関する技術は，本研究の直接の目的ではないが，分析対象となるタグの抽出や，タグ同士の関係性を表す指標として参考にする．

3. ハッシュタグ組織化モデル

ハッシュタグの組織化は，同じ意味のハッシュタグが並立している状況や，一つのタグにいくつかの異なる話題が混同している状況を前提とし，それらの状況を緩和するかたちで起こると考えられる．そこで本研究ではハッシュタグの組織化を以下の二つのモデルで捉える．

一つは，並立している複数のハッシュタグがひとつのハッシュタグにまとまる「収束」である．たとえば「紅白歌合戦」に関するハッシュタグは「#kouhaku」，「#kohaku」，「#nhk_kouhaku」など複数考えられるが，この中から一つのハッシュタグに集中して使われるようになることが収束である．

もう一つは，一つのハッシュタグ内に複数混在している話題に個別のハッシュタグが与えられる「分化」である．図 **1** のように，地震に関するハッシュタグ「#jishin」を使って，安否確認や地域別の情報などさまざまなツイートが投稿される状況が考えられる．ここから安否確認に関するハッシュタグ「#anpi」や，原発や地域ごとのハッシュタグが提案され，使い分けられていくのが分化である．

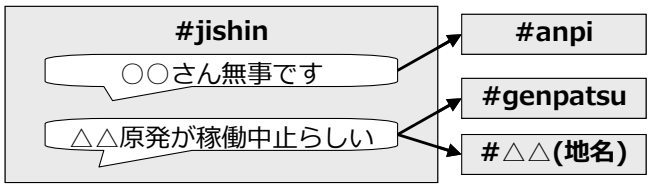

図 1 分化のイメージ

以下で収束と分化について事例を用いてより具体的に解説する．以下の二つの事例は，予備的な分析のために，組織化が起こっていると期待できるハッシュタグを人為的に選んだものである．

3.1. 収束の事例:「#NHK 紅白」

収束の事例として，第 62 回 NHK 紅白歌合戦[†]に関するハッシュタグについて述べる(以下,「紅白」と記す)．「紅白」には実況を目的としたハッシュタグが表記揺れによって複数並立して使われていた．特にツイート数の多かった並立ハッシュタグ 6 種類の時間帯ごとのツイート数を表 **1** に示す．たとえば「#NHK 紅白」を含むツイートは 19 時台（19 時 0 分から 19 時 59 分 59 秒）に 4447 件ツイートされている．「総ツイート数」とは，並立ハッシュタグ群のうち少なくとも 1 種類を含むツイートの数である．

表 1 紅白関連ハッシュタグのツイート数

時間帯	19 時	20 時	21 時	22 時	23 時
総ツイート数	7765	14797	7435	11109	8602
#NHK 紅白	4447	8534	4451	7432	6151
#紅白	1024	1470	1111	1115	758
#kouhaku	852	1871	580	897	611
#nhk_kouhaku	572	1559	561	743	598
#紅白歌合戦	822	1525	573	1078	494
#kohaku	223	218	165	212	167

ハッシュタグの収束の度合いを測る指標としてハッシュタグの「使用率」を定義する．使用率とは，ある並立ハッシュタグ群の総ツイート数に対する，あるハッシュタグを含むツイート数の割合である．たとえば「#紅白」の 20 時台の使用率は 1470÷14797≒9.9%となる（表 **1** より）．ハッシュタグの収束が進むということは，ある並立ハッシュタグのうち 1 つだけ使用率が高まり，その他ハッシュタグの使用率が低くなることであ、るといえる．

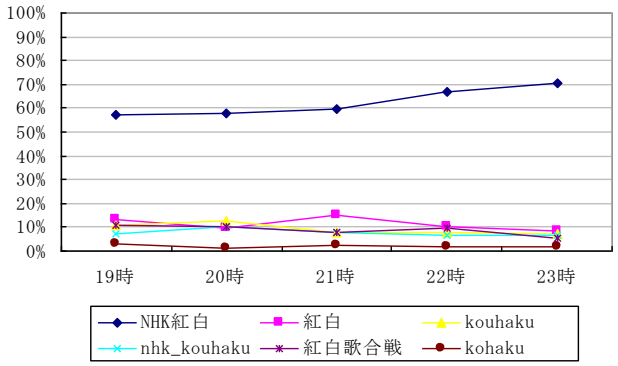

図 2 「紅白」関連ハッシュタグ使用率の変化

図 **2** は「紅白」関連ハッシュタグの使用率の変化を示したものである．約 4 時間の放送時間のあいだに「#NHK 紅白」の使用率が 57%から 70%まで上昇している．その他ハッシュタグの使用率はほぼ横ばいである．「紅白」関連ハッシュタグは「#NHK 紅白」が初めからよく使われ，さらに普及が進んだ一方，その他ハッシュタグもある程度継続して使われていたとわかる．

[†] 2011 年 12 月 31 日 19 時 15 分～23 時 45 分放送

なお,「#NHK 紅白」の使用率が最大なのは番組の公式アカウント‡がこれを使用していたことが主な理由と思われる.

3.2. 分化の事例:「#seiji」と「#TPP」

分化の事例として，政治全般に関するハッシュタグ「#seiji」と，環太平洋戦略的経済連携協定に関するハッシュタグ「#TPP」との関係について述べる.

分化は，様々な話題を包括するハッシュタグのなかで特定の話題が盛り上がり始めたとき，その話題に特化したハッシュタグが提案，使用されることで進むと考えられる．つまり，一般的・抽象的なハッシュタグほど分化が生じやすい.「#seiji」はそのような一般的なハッシュタグの一つといえる.

分化の度合いを計る指標として，ここではツイート数の変化と,ハッシュタグの共起率の変化に注目する.

2010 年 11 月から 2012 年 8 月における「#seiji」と「#TPP」を含むツイート約 240 万件について，ツイート数と共起率の変化を図 **3** に示す.「TPP / seiji」は「#seiji」を含むツイート数に対する「#TPP」を含むツイート数の割合である.「seiji -> TPP」は「#seiji」を含むツイートのうち「#TPP」を同時に含んでいる確率であり，「TPP -> seiji」はその逆である.

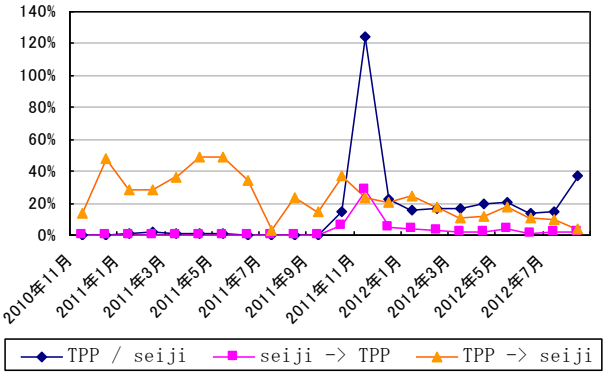

図 3 「#seiji」と「#TPP」におけるツイート数と共起率の変化

一般に，階層関係にあるタグについて，一般的なタグほど使用頻度が高く，特殊なタグは一般的なタグと共起する確率が高いと考えられる．この事例の場合，「#seiji」は「#TPP」よりもよく使われ（「TPP / seiji」が 100%以下で小さく），「#TPP」を含むツイートは同時に「#seiji」を含むことが多い（「TPP -> seiji」が大きい）.この性質を利用してタグの階層関係を抽出するのが ISR 手法[2]である．現に 2010 年 12 月から 2011 年 6 月頃まではそのような傾向が読み取れる.

しかし 2011 年 10 月頃から大きな変化がみられる.「#TPP」を含むツイートが急増し，2011 年 11 月には「#seiji」を含むツイート数を追い越している(TPP / seiji = 125%). また 2011 年 10 月からそれ以降にかけて「TPP -> seiji」は徐々に減少している．これは ISR 手法の前提に逆行するような興味深い現象だといえる.なお,この変化は 2011 年 11 月 11 日の野田首相(当時)の TPP 参加表明による影響が大きいと思われる.

ハッシュタグを利用するとき，似ているハッシュタグや関連するハッシュタグを併用することは一般的におこなわれる§. しかしハッシュタグを併用しすぎるとツイートが見づらくなったり，各ハッシュタグの検索結果をむやみに増えたりしてしまう問題がある．先述の Twitter 公式ブログ記事においても 1 ツイートあたりのハッシュタグ数を 2 つ以下にするよう助言されている*. このことから，ユーザはわかりやすさや利便性を保てるように併用ハッシュタグ数を調節するのではないかと期待できる.「TPP -> seiji」が小さくなったのは，TPP がもともと政治に関心の強いユーザに限った話題だった段階から，多くのユーザが関心をもつ重大な話題となる段階に移ったことを反映していると推測できる．もし急増した「#TPP」を含むツイートすべてに「#seiji」が共起していたら,「#seiji」検索結果の多くが TPP の話題で占められる．しかし併用を控えることによって，TPP の話題は「#TPP」を使い，それ以外の幅広い政治の話題はいままでどおり「#seiji」を使う，というように使い分けられる．このようにツイート数や共起率の変化は，階層関係の機械的な抽出だけでなく，ユーザがどのようにタグを使い分けているかを推測する指標としても利用できると考えられる.

4. 震災関連ハッシュタグの分析

本稿では[1]を参考に震災関連ハッシュタグを対象に組織化の分析を試みる．ハッシュタグ全般の組織化を分析するには幅広い分析対象が必要だが，すでに組織化の傾向が確認されている対象に絞ることでまずは効率的に分析することを目指す．また短期間に新しいハッシュタグが大量に模索され，ツイート数も膨大であったことから，特徴的な結果が得られると期待できる.

4.1. 分析対象

収束の分析は，同義あるいは類似しているハッシュタグ群が対象になる．分化の分析は，包括的なハッシュタグと，それに属する特殊なハッシュタグ，すなわち階層関係にあるハッシュタグ群が対象となる．[1]で分析された 77 種類のハッシュタグのうち,収束と分

‡ https://twitter.com/nhk_kouhaku

§ 2011 年 12 月 31 日 23 時 00 分から翌日 01 時 00 分までの「ハッシュタグを含むツイート」約 11 万件を対象に集計したところ，1 ツイートあたり平均 1.45 個のハッシュタグがつけられていた.

化の対象になると解釈できるハッシュタグ群を人為的に選出した．分析期間は，組織化がもっとも活発に進むと思われる初期段階として，2011 年 3 月 11 日から 31 日を対象とした．

表 **2** に収束の分析対象とする並立ハッシュタグ群を記す．たとえば，同じ「地震」を意味する 4 つのハッシュタグ「#earthquake」，「#eqjp」，「#jishin」，「#jisin」が収束するかどうかを分析する．

表 2 収束の分析対象となる並立ハッシュタグ群

地震・震災	#earthquake #eqjp #jishin #jisin
原発・放射線	#genpatsu #genpatu #houshasen
津波	#tunami #tsunami
医療	#311care #jishin_kusuri #311ER #ptsd_jp
募金	#bokin #gienkin #kifu
救え	#save_touhoku #savejapan #HelpJapan #save_japan

表 **3** に分化の分析対象とする階層ハッシュタグ群を記す．「#jishin」を，下に挙げた 5 つのハッシュタグを包括するものとし，たとえば「#jishin」から「#311care」が分化するかどうかを分析する．

表 3 分化の分析対象となる階層ハッシュタグ群

#jishin（地震）	
	#311care（医療）
	#genpatsu（原発）
	#tsunami（津波）
	#anpi（安否確認）
	#hinan（避難情報）

データの取得には Twitter API の一つである Twitter Search API を用いた．震災当時によく使われていたハッシュタグを優先的に，API 制限の範囲内で定期的に収集していたものであり，震災関連のすべてのハッシュタグつきデータを収集できている保証はない．規模の目安として，2011 年 3 月 11 日から 31 日に収集した「#jishin」ハッシュタグを含むツイート数は約 240 万件である（リツイートを含む）．

4.2. 分析方法

収束については，表 **2** の並立ハッシュタグ群を対象に，各ハッシュタグの使用率（3.1 項）の変化を計算する．使用率が最大のハッシュタグについての，使用率の高さと上昇の度合い，およびその他ハッシュタグの使用率の低さと減少の度合いから，収束の度合いを判定する．

分化については，表 **3** の「#jishin」（親ハッシュタグ）とその他 5 つのハッシュタグ（子ハッシュタグ）との関係それぞれを対象に，ツイート数と共起率（3.2 項）の変化を計算する．親ハッシュタグに対する子ハッシュタグのツイート数の増加や，子ハッシュタグにおける親ハッシュタグの共起率の減少が，分化の度合いを計る指標となる．

収束についても分化についても，このくらいの変化があれば組織化が起きているとみなす，という判定基準はいまのところ定められていない．今後，分析対象を増やしながらどのような傾向やパターンがあるかを踏まえたうえで指標や基準を洗練していく必要がある．

4.3. 収束の分析結果

図 **4** に「地震・震災」群のハッシュタグ使用率を示す．震災直後，4 つのハッシュタグのうち「#jishin」の使用率が 72%でもっとも高い．「#jisin」も 35%，その他 2 つも 20%程度使われている．3 月 12 日，13 日にかけて「#jishin」の使用率が上昇し，それ以後も 90%以上を維持している．同時に「#jisin」以外の使用率は減少している．「地震・震災」群における最初の 2 日間の変化は分析対象のなかで特に顕著だった．

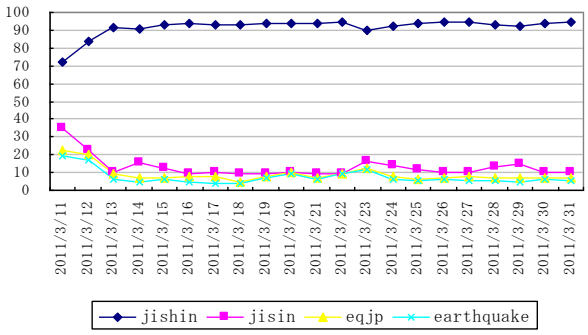

図 4 「地震・震災」群のハッシュタグ使用率

「原発・放射線」，「津波」，「医療」群については，震災直後から，もっとも使用率の高い 1 つのハッシュタグが 90%以上の使用率を維持している（それぞれ「#genpatsu」，「#tsunami」，「#311care」）．その他のハッシュタグについてはほぼ 10%以下の使用率だったが，「#genpatsu」と並立する「#genpatu」については使用率が 40%や 25%と高い日があった．

図 **5** に「募金」群のハッシュタグ使用率の変化を示す．震災直後は「#bokin」がほぼ 100%の使用率だったが，3 月 12 日から 13 日にかけて「#gienkin」の使用率が上昇している．「#bokin」の使用率が一貫して最大ではあるが，上昇と下降を不規則に繰り返している．

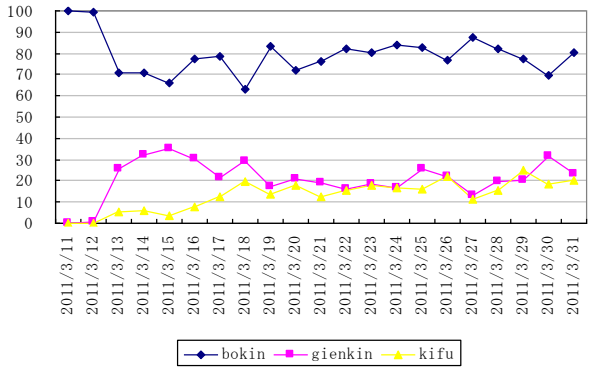

図 5 「募金」群のハッシュタグ使用率

図 **6** に「救え」群のハッシュタグ使用率の変化を示す．震災後数日間は使用率が最大のハッシュタグが何度か入れ替わっている．その後は「#save_touhoku」が最大となるが，月末には「#save_japan」との差が小さくなっている．

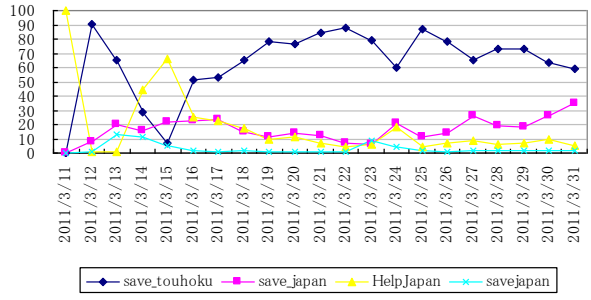

図 6 「救え」群のハッシュタグ使用率

4.4. 分化の分析結果

図 **7** に「地震」と「避難情報」の関係の変化を，図 **8** に「地震」と「安否確認」の関係の変化を記す．両方とも震災直後に子ハッシュタグにおける親ハッシュタグとの共起率が高い．2 日かけて共起率が下降し，4 日目ですこし上昇，それ以降で下降，というように傾向が似ている．「安否確認」のツイート数(anpi / jishin)は月末にかけてすこしずつ減少しているが，「避難情報」(hinan / jishin)はほぼ横ばいである．

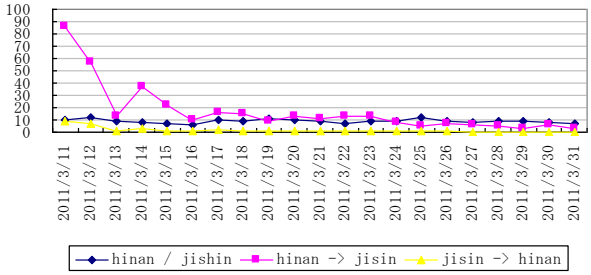

図 7 「地震」と「避難情報」の関係

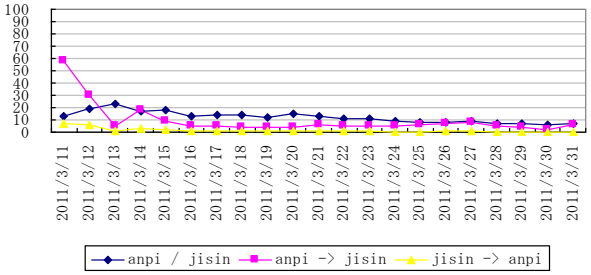

図 8 「地震」と「安否確認」の関係

図 **9** に「地震」と「津波」の関係の変化を示す．「津波」のツイート数(tsunami / jishin)は震災当日に 26%の割合を占めていたが，2 日かけて 6%に減少している．その後しばらく親ハッシュタグとの共起率(tsunami -> jishin)が 40%以上を保ち，3 月下旬は減少している．

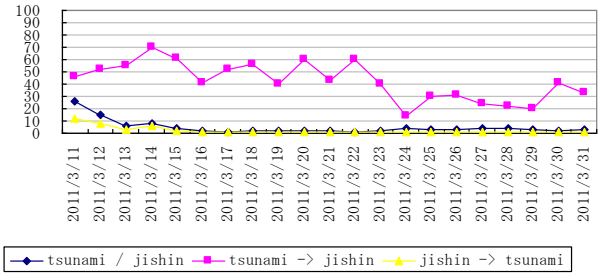

図 9 「地震」と「津波」の関係

図 **10** に「地震」と「原発」の関係の変化を示す．「原発」のツイート数(genpatsu / jishin)は震災直後の数日は低く，その後徐々に増加している．逆に親ハッシュタグ(genpatsu -> anpi)との共起率は徐々に減少している．3 月 22 日から 23 日にかけてツイート数が倍以上に増加している．

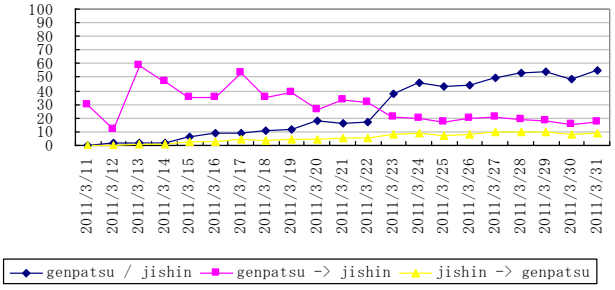

図 10 「地震」と「原発」の関係

図 **11** に「地震」と「医療」の関係の変化を記す．「医療」のツイート数(311care / jishin)は震災の翌日に増加し，その後も横ばいである．親ハッシュタグとの共起率(311care -> jishin)は緩やかに上下している．

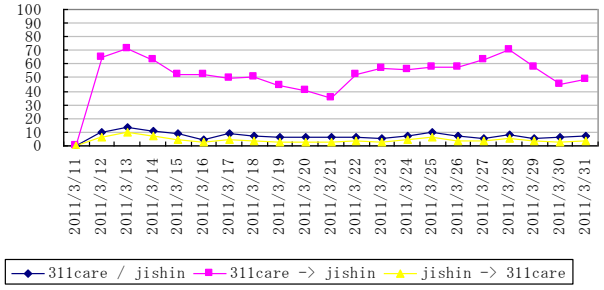

図 11 「地震」と「医療」の関係

4.5. 結果の考察

収束については，6 つのうち 4 つのハッシュタグ群で使用率 90%以上を維持するハッシュタグがみられた．とくに「#jishin」の使用率は震災直後に一気に上昇していることから，ユーザは表記揺れで並立しているハッシュタグ群から 1 つを選んで集中的に使おうとする傾向が読み取れる．使用率の高かった「#jishin」，「#311care」，「#tsunami」は Twitter 公式ブログ記事*で推奨されていたが，「#genpatsu」や「#bokin」は記載されていなかった．記事は震災の翌日に投稿されており，その時点では原発や募金の重要性が認識されていなかったためだと思われる．

また地域別ハッシュタグについては，本稿の「救え」群で対象にした「日本」や「東北」などの広い範囲ではなく，「#save_miyagi」などの都道府県レベルで記事には記載されていた．さらに「save」と「help」のような日本人にはニュアンスの違いがわかりにくい表記揺れや，地域の粒度の使い分けの難しさによって「救え」群は収束に混乱をきたしていたと思われる．

分化については，「医療」を除いた 4 つの群において，子ハッシュタグにおける親ハッシュタグとの共起率が減少していく傾向が共通していた．とくに震災直後の 2 日間で，「安否確認」は 58%から 5%へ，「避難情報」は 86%から 13%へと，極端な変化がみられた．初期段階にハッシュタグを併用することで「#hinan」や「#anpi」が震災に関するものであると強調され，多くのユーザがハッシュタグの意味や使い方を理解できるようになるだろう．その後は目的に特化して単独でハッシュタグが使われるようになったと読み取れる．

「原発」については，「#seiji」と「#TPP」（3.2 項）と共通するようなツイート数の増加と共起率の減少の同時発生がみられた．ここから「原発」が「地震」全般に匹敵する重大な関心事として認識されたとわかる．

5. おわりに

フォークソノミーにおけるタグの並立や混同という問題を，Twitter のハッシュタグではユーザが時間をかけて使い分けていくことで緩和していく，ハッシュタグの自律的な組織化に注目し，震災関連ハッシュタグを対象に分析した．複数の並立ハッシュタグが 1 つのハッシュタグに集約していく「収束」や，包括的なハッシュタグのなかで生じた話題が別のハッシュタグに移ってツイートされていく「分化」という現象を確認した．収束においては，Twitter 公式ブログのような信頼できる情報原からのよびかけが有効だと思われた．分化の事例では，話題の発生直後はハッシュタグの積極的な併用によって情報共有を促し，徐々に目的に特化して使い分けられていく傾向がみられた．

このようにいくつかの組織化の事例とその傾向が読み取れたが，なぜそのような傾向があるのかという要因については今後も検討が必要である．たとえば，ユーザが自主的に推奨ハッシュタグをよびかけるようなツイートや，リツイートされるツイートの傾向などは影響が大きいと思われる．分化について詳しく分析するには，共起率だけでなくキーワードの変化にも着目したい．本稿では一部の事例のみを対象としたが，幅広い事例を積み重ね，このような組織化が一般にどのように起こるかを明らかにしていきたい．

「#seiji」と「#TPP」，あるいは「#jishin」と「#genpatsu」にみられるように，どのハッシュタグが盛り上がるかというのはユーザの関心の集まりとその変化を表している．適切にハッシュタグを使い分けることで関心を深く共有できる場が用意され，ユーザ同士のより有意義な交流が可能になるだろう．ハッシュタグの自律的な組織化とは，そのような場をユーザみずからがつくろうという姿勢の表れであると思う．

参 考 文 献

[1] 村井源，“東日本大震災後の Twitter 利用傾向 -震災関連ハッシュタグの計量的分析-”，情報知識学会誌 22(2), 97-106, (2012)

[2] Rui Li, Shenghua Bao, Ben Fei, Zhong Su, Yong Yu, “Towards Effective Browsing of Large Scale Social Annotations”, WWW2007, pp.943-952, (2007)

[3] 谷田川将之，永森光晴，杉本重雄，“Twitter ハッシュタグの構造化に関する研究”，情報処理学会第 73 回全国大会講演論文集 2011(1), 693-695, (2011)

[4] 多田亮平，湯本高行，新居学，高橋豊，“ソーシャルタグの表記と使用傾向に基づく集約”，DEIM Forum 2012, F2-2, (2012)

[5] 丹羽智史，土肥拓生，本位田真，“Folksonomy の 3 部グラフ構造を利用したタグクラスタリング”，人工知能学会研究会資料，SIG-SWO-A602-07, (2006)

[6] 村上直至，伊東栄典，“動画投稿サイトで付与された動画タグの階層化”，情報処理学会研究報告，2010-MPS-81, 17, (2010)

[7] 杉本徹，五十嵐 幹，“Folksonomy におけるタグ語の意味的階層関係の抽出”，情報科学技術フォーラム講演論文集, 7(2), 253-254, (2008)