1F2-OS8-4

Twitterの情報伝播ネットワークの分析

Analysis of Information Diffusion Network on Twitter

風間 一洋　今田 美幸　柏木 啓一郎
Kazuhiro KAZAMA　Miyuki IMADA　Keiichiro KASHIWAGI

日本電信電話 (株) NTT 未来ねっと研究所
NTT Network Innovation Laboratories, Nippon Telegraph and Telephone Corporation

This paper analyzed information diffusion networks, which are created by communication on Twitter, to find interesting or useful discussion around a user. We classify functions related to information diffusion and propose the method to extract information diffusion networks around a user. Furthermore, we extract and analyze information diffusion networks within 3 hops from a specified user. In the result, we show the characteristics that is useful for finding interesting or useful discussion.

1. はじめに

Twitterは，現在何をしているかなどのつぶやき（tweet）を投稿するサービスである．発言が140文字に制限される反面，リアルタイムに相手に伝わることから，既存のニュースサイトよりも早く情報を伝達・収集できる手段やある話題についてリアルタイムに議論するための手段として活用され，Twitterにおける情報発信・交換はブログ以上に盛んになりつつある．

ただし，フォロー数を増やせば増やすほどツイート数が膨大になり，すべてに目を通すことが困難になるために，情報入手範囲の拡大と妥当な情報量の両立はTwitterにおけるジレンマである．フォローしているユーザでもフォローしていない相手に対する返信はタイムラインに表示されないこともあり，Twitterのフォロー関係における自分の近傍から興味深い議論の発見を支援する仕組みが望まれる．

本稿では，Twitter上で何が起こっているのかを知るための基礎技術を確立するために，情報伝播現象を分析する．まず最初に情報伝播に関係したTwitterの機能を調査し，それらを元に情報伝播経路を抽出する手法について述べる．さらに，抽出した情報伝播経路の特徴を分析して，情報伝播ネットワークの性質や，興味深い議論を発見するために有用と思われる特徴について述べる．

2. 情報伝播ネットワーク

2.1 Twitterにおける情報伝播

Twitterでは，あるユーザのツイートがフォロー関係やハッシュタグ[*1]の検索結果を通じて伝播して他のユーザに伝わり，それに触発されて書いたツイートがさらに伝播する現象が連鎖的に発生する．このような情報が伝播した経路は，ネットワーク構造として抽出できる．

このネットワークは，ユーザ間のネットワークと情報間のネットワークの2種類に分類できる．前者は一種の社会ネットワークであり，単なるフォロー関係ネットワークと比較すると，ネットワークの同質選択制が高くなり，面識があったり，興味が非常に近いユーザ間の割合が高いと考えられる．後者は，前者を時間軸に対して展開・分解して得られるネットワーク構造と考えることができ，一般的に放射状に広がる形状を持つ．本稿では，後者の情報間のネットワーク構造について分析する．

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

*1 “#” を先頭につけた文字列で，話題やジャンルの分類に用いる

2.2 情報伝播の判定

フォロー関係にあるユーザ間を情報が伝播したかどうかは，以下の2種類の方法で判定されている．

1. 潜在的に伝播可能の複数の経路の中から，実際に伝播した経路を推定する
2. ユーザ間の明示的な情報交換を経路と見なす

前者の手法は，例えば Adar らがブログに適用しており [Adar 05]，明示的に伝播関係が示されていない場合も扱えるので，より現実に近い伝播経路を推定できる可能性がある．ただし，Twitterの場合はテキスト長が短いことからテキスト類似度だけで判断しにくく，またバースト性が強いTwitterでは単純に直近のツイートと関係づけることもできず，ブログよりも判定が難しい．また，フォロー関係による情報伝播とハッシュタグの検索結果による情報伝播を同様には扱えない．

後者の手法では，実際に明示的な情報交換をおこなったユーザ間だけに情報の伝播があったと見なす [風間 10]．実際には，ツイートのそのユーザ自身の発言部分内に存在するユーザ名やURL，Twitter APIが返すユーザやツイートの関係などで判定する．ただし，この手法では，実際に情報の伝播が起こっていても明示的な関係がない場合には判断ができず，前者の手法よりも得られる情報伝播経路の規模が小さくなる問題がある．本稿では，後者の手法を利用する．

2.3 情報伝播の種類

Twitterでは，以下の4種類の情報伝播が観測できる．ダイレクトメッセージも提供されているが，メッセージの送信者・受信者以外は見ることはできないので本稿では除外する．

1. 返信
2. リツイート
3. 非公式リツイート
4. URL

1は，別のユーザのツイートに対する返信である．この例を図1に示す．テキスト中に “@ユーザ名” 又は “.@ユーザ名” を含み，`in_reply_to_status_id`，`in_reply_to_user_id`，`in_reply_to_screen_name` により元のユーザ及びツイートが

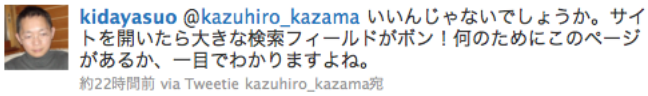

図 1: 返信の例

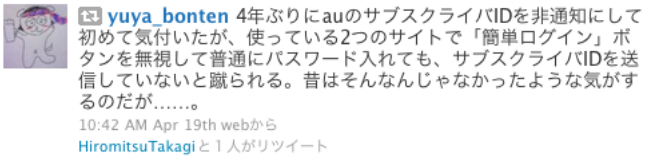

図 2: リツイートの例

特定できる．“@ユーザ名” と “.@ユーザ名” の違いは伝播範囲であり，前者は返信する相手もフォローしているユーザだけだが，後者はすべてのユーザに表示される．

2 のリツイート（retweet）は，別の人のツイートを自分のフォロワーにも知らせる時に用いる機能である [Williams 09]．この例を，図 2 に示す．`retweeted_status` により引用元のユーザ及びツイートが特定できる．

公式なリツイートは 2009 年 11 月に英語インタフェースにおいて徐々に導入され，日本語インタフェースでも 2010 年 01 月 22 日に利用できるようになったが，その前から 3 の非公式なリツイートがユーザから自発的に生まれ，一部のサードパーティの Twitter クライアントによりサポートされていた．これを本稿では非公式リツイートと呼ぶ．この例を図 3 に示す．非公式であるために API によるサポートはない．フォーマットも “RT @引用元ユーザ名: 引用元ツイート”，“コメント RT @引用元ユーザ名: 引用元ツイート”，“コメント QT @引用元ユーザ名: 引用元ツイート” など複数存在する．公式なリツイートの利用率は増えつつあるが，コメントが記述できない，クライアントが未サポートという理由から，いまだに非公式リツイートも多く使われている．

4 は，メッセージで同じ URL について言及していた場合であり，その場合はその URL からツイートに対して情報が伝播したと見なすことができる．この例を図 4 に示す．ただし，URL 短縮サービスが用いられる場合は，一旦元の URL に復号化する必要がある．

3. 情報伝播経路の抽出手法

Twitter の情報にアクセスするために提供されている Twitter API[Twitter] を用いて情報伝播経路を抽出する方法について述べる..

3.1 情報伝播経路の抽出

情報伝播経路のネットワーク構造を抽出する大まかな手順は，以下の通りである．

1. 各ユーザのタイムラインを取得する．

図 3: 非公式リツイートの例

図 4: URL の例

2. 返信，リツイート，非公式リツイートで関連するツイートをグループ化する．
3. さらに同一の URL を参照しているグループを統合する．
4. 各グループの有向ネットワーク構造を出力する．

3.2 非公式リツイートの処理

非公式リツイートは公式 API ではサポートされないので，独自にツイートを解析する必要がある．例えば図 3 は，“ユーザコメント RT @引用元ユーザ名: 引用元ツイート” というフォーマットであり，解析の結果ユーザコメントと引用元ツィートのテキストと引用元ユーザ名を得ることができる．ただし，いくつかの問題が存在する．

一つ目は，非公式ツイートには複数のフォーマットがあり，さらに自然発生的に生まれたために，ユーザ名の直後の “:” の有無などの細かい違いや論理的な曖昧さがあることである．例えば図 3 は，元の発言が “QT”，“RT” と異なるフォーマットで 2 回リツイートされている．このように包含構造を持つ場合は，異なるフォーマットが混在していても正しく解析しなければならない．本稿では，異なるフォーマットを統一的に解析するのが困難だったので，本稿では複数のフォーマットを並行に解析し，一番外側の解析結果を使用した．ただし，ユーザが手入力する場合には文法エラーが発生している確率が高く，ユーザ名の直後の “:” や半角スペースの省略，全角文字の誤用，複数ユーザの指定（本来は単一ユーザ）などの問題が頻繁に発生しており，完全に対応するのは難しい．

二つ目は，非公式リツイートでは，返信やリツイートのような引用元ツイートの情報が得られないことである．さらに，非公式リツイートであることや引用元ユーザ名などのメタ情報はテキスト中に表記されるために，非公式リツイートが繰り返されるにつれて引用元のテキストが切り捨てられてしまう．本稿では，すべてのツイートのテキストを前方一致で検索して，引用元を発見した．ただし，リツイートされたテキストが短い場合は誤認識する確率が高いので除外している．

なお，非公式リツイートでは，リツイートしたユーザに関する情報はコメント部分だけである．そこで，非公式リツイートから URL を抽出する場合は，コメント部分に存在する URL だけを対象とし，引用元ツイートの部分に存在しても無視している．

3.3 情報伝播ネットワークの例

以上の手法で多くの情報伝播ネットワークが抽出できるが，その一例を図 5 に示す．楕円形で示したノードはツイートであり，矩形で示したノードは URL である．また，直線，破線，点線の矢印は，それぞれ返信，非公式リツイート，URL による情報の伝播を示し，矢印の方向は生成時刻が新しい方向，つまり情報伝播の方向を表す．

4. 情報伝播ネットワークの分析

4.1 データセット

Twetter4J[Yamamoto 07] を用いて実装したクローラーで，あるユーザから 3 ホップ以内のユーザの 2010 年 4 月 1 日から 2010 年 4 月 15 日の間のツイートを収集した．なお，返信の

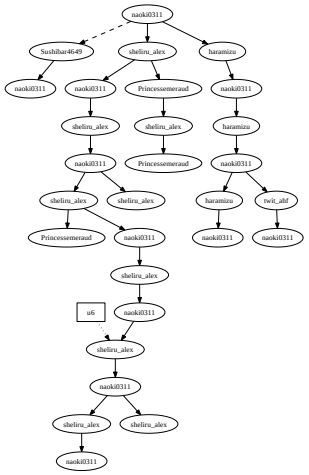

図 5: 情報伝播ネットワークの例

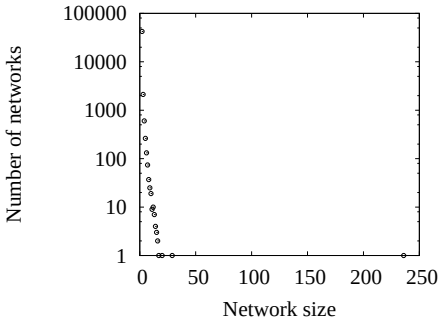

図 6: サイズの分布

場合は返信元情報が得られるのでさらに 1 ホップ先まで処理対象になるが，非公式リツイートの場合には実際に引用元ユーザのタイムラインを取得しないと判定できない．

ただし，現在の Twitter API の仕様では，後方互換性を保つために XML 又は JSON を用いるとユーザタイムラインにリツイートを表示しない制限があり，それを補うために追加された retweeted_by_me，retweeted_to_me，retweets_of_me では任意のユーザのリツイート情報を収集できないので，今回はリツイートは収集できなかった [Twitter]．

収集したユーザ数は 1,752 人であり，1 人のユーザを起点として 1 ホップが 6 人（0.3%），2 ホップが 92 人（5%），3 ホップが 1,653 人（94%）であった．

全ツイート数は 185,065 で，返信は 50,538（27%），非公式リツイートは 14,224（8%），URL を含むツイートは 33,646（18%）であった．

抽出した全 URL 数は 34,853 で，短縮 URL は 16,263（47%）であった．ただし，これは出現頻度が高い七つの短縮 URL サービスについて調べたものであり，実際にはもう少し割合が多い可能性がある．

4.2 サイズとユーザ数の分析

本稿で述べた手法では，複数の情報伝播ネットワークが得られる．そこで，ツイート数が 2 以上の抽出された情報伝播ネットワークに対して，x 軸にネットワークに含まれるツイート数を，y 軸にそのツイート数のネットワーク数をプロットしたグラフを，図 6 に示す．この分布は直線状になるために，指数分布に従うと推測される．平均ツイート数は 2.14 であるが，これは単一の返信の割合が非常に多いからである．

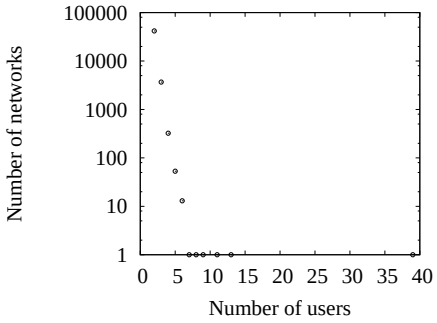

図 7: ユーザ数の分布

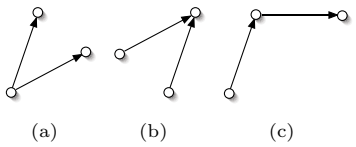

図 8: 情報伝播ネットワークの基本構造

次に，x 軸にネットワークに含まれるユーザ数を，y 軸にそのユーザ数のネットワーク数をプロットしたグラフを，図 7 に示す．平均ユーザ数は 2.1 である．

情報伝播ネットワークのサイズは議論の盛り上がりを示しており，サイズが大きいネットワークほど議論が活発におこなわれていると推測できる．ただし，サイズに比べてユーザ数がかなり減少していることから，一つの情報伝播ネットワーク中に同じユーザのツイートが繰り返し出てくることが多いことがわかる．

なお，ツイート数が 236 と極端に多い情報伝播ネットワークが抽出されているが，このユーザ数は 39 人にすぎない．この情報伝播ネットワークを調べると，同一の URL を含む大量のリツイートをしている 2 人のユーザが発見され，一種のスパムであると思われる．

4.3 基本構造数の分析

Twitter はツイートの削除だけで変更は許されないために，ツイート間の因果関係を考慮すると情報伝播ネットワークは無閉路有向グラフになり，図 8 に示すような図 8(a) は情報の拡散，図 8(b) は情報の集約，図 8(c) は情報の転送という 3 種類の有向 2 エッジ部分グラフで構成される．それぞれを情報拡散構造，情報集約構造，情報転送構造と呼ぶ．同様な方法で抽出したブログの情報伝播ネットワークと比較すると，Twitter の情報伝播ネットワークは図 5 の例と同様に，ほとんどが情報拡散構造と情報転送構造で構成されるという特徴がある．これは返信元とリツイート元は単一であるという Twitter の制約によるものであり，情報集約構造が生じるのは同時に URL を記述した場合に限られる．

Twitter 上で特に活発に議論がおこなわれている箇所に着目するためにサイズ 5 以上の情報伝播ネットワークを抽出して，Twitter で特徴的な情報拡散構造と情報転送構造を調査した．情報伝播ネットワークの情報拡散構造数と情報転送構造数の分布をそれぞれ図 9 と図 10 に示す． 情報拡散構造が 0 のネットワークは 278 個で，これは情報転送が直線状におこなわれているネットワークである．情報転送構造が 0 のネットワークは 50 個で，これは 1 段階の情報拡散のネットワークである．さ

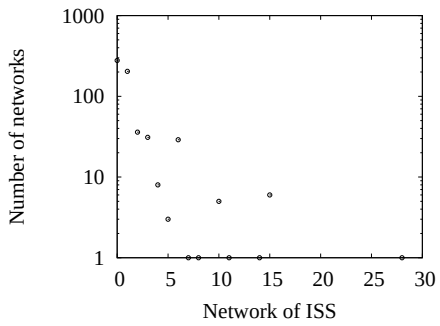

図 9: 情報拡散構造数の分布

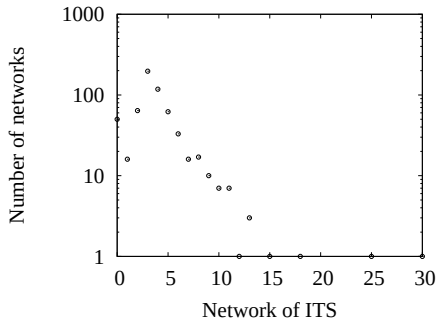

図 10: 情報転送構造数の分布

らに情報拡散構造が 1 のネットワークが 204 個，情報転送構造が 1 のネットワークが 16 個であることを考慮すると，Twitter は基本的に情報転送が連鎖的におこなわれているネットワーク構造であることがわかる．情報拡散構造は 0 を最大として減少する分布であるが，情報転送構造は 3 をピークとする分布であった．これは，サイズ 5 の直線状伝播のネットワークの情報拡散構造数は 3 であり，これが特に多いことが影響していると思われる．このために，ブログと比較すると Twitter の方が情報伝播経路長が長くなる傾向がある [風間 10]．これはコメント，トラックバックを含めた複数の手段を使い分けねばならないブログに比べると，統一的で簡単な方法で返信やリツイートがおこなえるからであると推測する．

なお，サイズが大きい Twitter の情報伝播ネットワークは，図 5 の例でもわかるように，一言で言えば多数の直線状構造とそれより少ない分岐で構成される傾向がある．ただし，直線状構造は長くなってもユーザ数は 2 人である場合が多い．長さは議論の活発さは示すが，議論の内容は個人的な内容である可能性も高いと考えられる．

また，広く分岐している部分でも，図 11 に示すように同一ユーザによる返信・非公式リツイートが繰り返されている場合も多い．この理由としては，140 文字という制限で 1 回で書ききれないために分割した，後から新たな情報を書き込んだなどが考えられる．

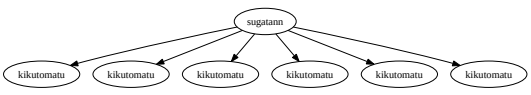

図 11: 同一ツイートに対する返信の繰り返しの例

4.4 考察

Twitter のフォロー関係における自分の近傍からの興味深い議論の発見という条件は，内容の適合性と注目度である．

内容の適合性は基本的に使用されている単語の出現頻度などを利用すると思われるが，その場合でも本稿のように情報伝播ネットワークとして抽出すれば，文字数制限のあり内容が関連しているツイートをまとめて扱うことができ有用である．

注目度は，一般的な Web 情報の場合は URL の引用数から，ツイート単体の場合はリツイート回数などから推測できるが，本稿で注目している複数のツイートで構成される議論の場合はより複雑である．そこで，議論の活発さを示すツイートの情報転送の連鎖が長く，多くの人の興味を引いていることを示す複数ユーザへの分岐構造の数と分岐数が多い情報伝播ネットワークが，注目されている議論だと考えられる．

なお，URL の引用は，始点ツイートで引用される場合と，図 5 のように返信で引用される場合に分類できる．前者はニュースを話題のきっかけとするような議論であるが，後者は質問に対して答える議論であることが多い．このような Q&A 的な議論の検出も有用であると考えられる．

5. おわりに

本稿では，Twitter から返信や非公式リツイート，本文中の URL を手がかりに情報伝播ネットワークを抽出する方法について述べて，さらに実際にあるユーザの近傍のツイート群から抽出した情報伝播ネットワークを分析し，有用と思われる

今後はユーザの近傍の興味深い議論の効率的な発見手法と，その議論の種類の自動判定をおこなう予定である．

参考文献

[Adar 05] Adar, E. and A.Adamic, L.: Tracking Information Epidemics in Blogspace, in *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 207–214 (2005)

[風間 10] 風間 一洋, 今田 美幸, 柏木 啓一郎：ブログ空間の情報伝播ネットワーク特性の定量化, 人工知能学会論文誌, Vol. 25, No. 3, pp. 404–409 (2010)

[Twitter] Twitter, : Twitter API Wiki, `http://apiwiki.twitter.com/`

[Williams 09] Williams, E.: Why Retweet Works the Way it Does, `http://evhead.com/2009/11/why-retweet-works-way-it-does.html` (2009)

[Yamamoto 07] Yamamoto, Y.: Twetter4J, `http://twitter4j.org/` (2007)