

令和元年度 卒業論文

Twitter アカウトを用いた拡散過程の特徴抽出

千葉大学工学部都市環境システム学科

16T0317W 井上陽向

指導教員：塩田茂雄

令和2年2月3日提出

目次

第1章 序論.....	1
1.1 研究背景.....	1
1.2 研究目的.....	1
1.3 既存研究.....	2
1.4 論文構成.....	3
第2章 Twitter データ	4
2.1 Twitter.....	4
2.2 用語.....	6
2.3 TwitterAPI.....	10
2.3.1 概要.....	10
2.3.2 OAuth 認証	10
2.3.3 パラメータとデータ構造.....	11
第3章 リツイート特徴.....	13
3.1 フォロワー特徴分析.....	13
3.1 フォロワー数とリツイート数関係.....	16
第4章 拡散経路分析.....	18
4.1 分析手法	18
4.1.1 フォロワー・フォロワーネットワーク.....	18
4.1.2 1次拡散者, 情報源不明, 非公開リツイートユーザーの定義	21
4.2 経路分析の結果.....	24
4.2.1 リツイート数と各割合の関係.....	24
4.2.2 フォロワー数と各割合の関係.....	29

第 5 章 結論.....	31
謝辭.....	32
参考文献.....	33

第 1 章

序論

1.1 研究背景

近年、スマートフォンの普及により、Twitter や Facebook などのソーシャルネットワークワーキングサービス（SNS）は人々の情報収集や情報発信、コミュニケーションの日常的なツールとなっている。なかでも Twitter の日本国内での月間アクティブユーザー数は約 4500 万人であり、人気のサービスあることがうかがえる。

SNS においては、日々、多様な情報が発信、拡散され、我々の生活に大きな影響を及ぼしている。リアルタイムで様々な情報が閲覧できるため、その即時性から、災害などの重大事象が発生した際には、貴重な情報源になることがある。近年、個人の情報収集やコミュニケーションの場だけでなく SNS が持つ情報拡散力を活用して、企業のマーケティングや地方自治体の広報活動の場として活用する動きがある。本研究にあたり、Twitter アカウントの運用を行う企業 A から、フォロワーの特徴分析の依頼を受けた。フォロワーの中でどのようなユーザーがリツイート（拡散）を行う傾向があるのか、何が要因となり、リツイート数が増加しているのかを明らかにすることは、企業など情報を広く拡散したいという要望のあるユーザーのマーケティングの活動や広報活動に重要であると考えられる。

1.2 研究目的

本研究の目的は、TwitterAPI を用いて、いくつかの企業や自治体、政治家の Twitter アカウントを対象に分析を行い、リツイート回数多いユーザーの特徴を明らかにすること、ツイートデータの拡散経路分析を行い、リツイート数増加の要因を明らかにすることにある。

1.3 既存研究

SNS の情報拡散において、どのユーザーが拡散において重要な影響を及ぼすのか、ネットワークの構造による拡散の影響があるのかなど様々な研究がなされている。以下にいくつかの情報拡散に関連する既存研究とその説明を行う。

Cheong らの研究[1]では、オーストラリア洪水の際に Twitter ユーザーが実際に投稿したツイートを対象に分析を行い、ネットワーク上での拡散における重要アカウントの特定手法とその結果を示している。小出らの研究[2]では、ネットワーク中から発信される情報を受け取り、その情報を多くのノードに伝搬する高影響力をもつノード（超媒介者）の検出方法を提案している。また、臼井ら[3]は、情報拡散ネットワークからいくつかの構造特徴を抽出し分析を行うことで、ネットワークの構造が情報拡散に与える影響を明らかにしている。那須野ら[4]は、Twitter 上での情報拡散とその影響に着目し、国政選挙での候補者のツイートを対象に分析を行い、いくつかの指標を用いて、当選者予測の手法の検討を行っている。

その他にも、南川らの研究では、複雑ネットワーク上の情報拡散に着目し数理モデルを用いて分析を行うことによって、ノード相関（隣接するノード状態の関係性）が情報拡散過程に大きく影響することを明らかにしている。[5]また、キーワード検索により収集した Twitter データをもとに分析を行うことで、非日常的な出来事に関するキーワードは、リツイートが大半を占めること、単位時間あたりのリツイート数変化は、ピークを迎えた後に減衰する定型パターンに従うこと、リツイート数とフォロワー数の相関は小さいことを明らかにしている。[6]

このように、SNS の情報拡散やネットワーク着目した研究はいくつか存在するが Twitter 上でのデータをもとに拡散経路を分析した研究はほとんどない。本研究では、拡散経路の分析において、リツイートユーザーの中で 1 次拡散者、情報源不明ユーザー、非公開ユーザーに着目し、リツイート数とそれぞれの関係を分析することによって、リツイート数増加要因を考察する。また、企業 A のアカウントを対象として、フォロワーの分析を行い、リツイート回数の多いユーザーの特徴を明らかにする。

1.4 論文構成

以下、本論文の構成内容を述べる。

第1章

本研究の研究背景、研究目的、既存研究、構成を述べる。

第2章

研究内容である Twitter の特徴、本稿で用いた基本的な用語の説明や、TwitterAPI の説明と分析に用いた TwitterAPI のパラメータとその構造について述べる。

第3章

企業 A を対象にしたフォロワー分析の結果とその特徴の考察、複数アカウントを対象にしたリツイート数とフォロワー数関係の分析結果について述べる。

第4章

前半に、拡散経路分析のためのフォロワー・フォロワーネットワーク作成手順の説明と、1次拡散者、非公開ユーザー、情報源不明ユーザーの定義を示す。後半に、複数のアカウントを対象に分析を行った結果を述べ、リツイート数増加の要因を考察する。

第5章

本研究の結論と今後の課題について述べる。

第 2 章

Twitter データ

2.1 Twitter

Twitter は、Facebook や Instagram, LINE にならぶ SNS (Social Networking Service) の 1 つである。スマートフォンの普及とともに、SNS 利用者数は急速に拡大してきている。総務省の平成 30 年通信利用動向調査の結果によると、SNS を活用している割合は、個人が 60.0%、企業が 36.7%となっている。図 2.1 に平成 30 年通信利用動向調査[7]の SNS の利用目的 (個人) のグラフを、図 2.2 に SNS の利用目的 (企業) のグラフを示す。個人の利用目的としては、「知人とのコミュニケーションのため」「知りたいことについて情報を探すため」「ひまつぶしのため」が上位を占めており、情報の収集やコミュニケーションツールとして、SNS が個人に浸透していることがわかる。企業の利用目的としては、「商品や催物の紹介、宣伝」「定期的な情報の提供」「会社案内、人材募集」が上位を占めている。

Twitter は、国内月間アクティブユーザー数は 4500 万人、世界全体ユーザー数 3 億 3500 万人と世界中で人気のサービスである。140 文字 (英数半角で 280 文字) を上限として、誰でも気軽にツイートする (つぶやく) ことができる。リアルタイムで様々な情報が閲覧できるため、情報に即時性があり、リツイートと呼ばれる機能によって情報の拡散力が高いという特徴がある。ユーザーは他のユーザーをフォローすることによって、自身のタイムライン上でフォローしたユーザーのツイートを閲覧することが可能になり、「リツイート」や「いいね」、「リプライ (返信)」機能によって、情報の拡散や他ユーザーとのコミュニケーションを行う。また、近年、「トレンド」と呼ばれる機能が実装され、世界で今何が話題であるかという情報が閲覧できるようになるなど、年々新たな機能の実装が行われている。この Twitter の拡散力の強さが注目され、自治体や企業、政治家の情報発信やマーケティングの場として利用する動きがある。このように Twitter は気軽に不特定多数の人物に情報の発信、拡散が行えるといった利点がある。しかし一方で、根拠のないデマ情報が拡散されたり、個人の発言が炎上するといった危険も存在する。

ソーシャルネットワーキングサービスの利用目的(個人)

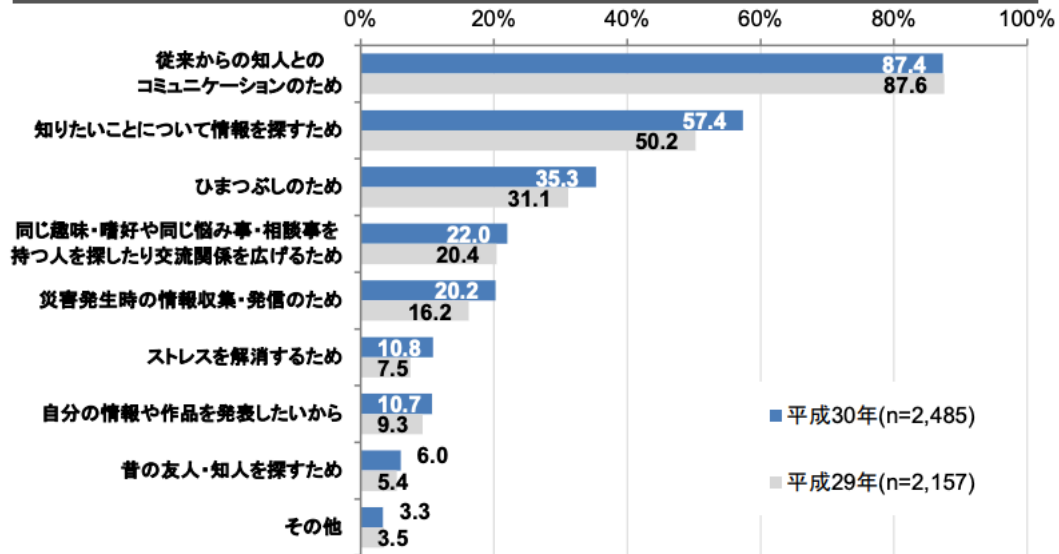


図 2.1 : SNS の利用目的 (個人) (平成 30 年, 平成 29 年)

ソーシャルメディアサービスの活用目的・用途(企業)

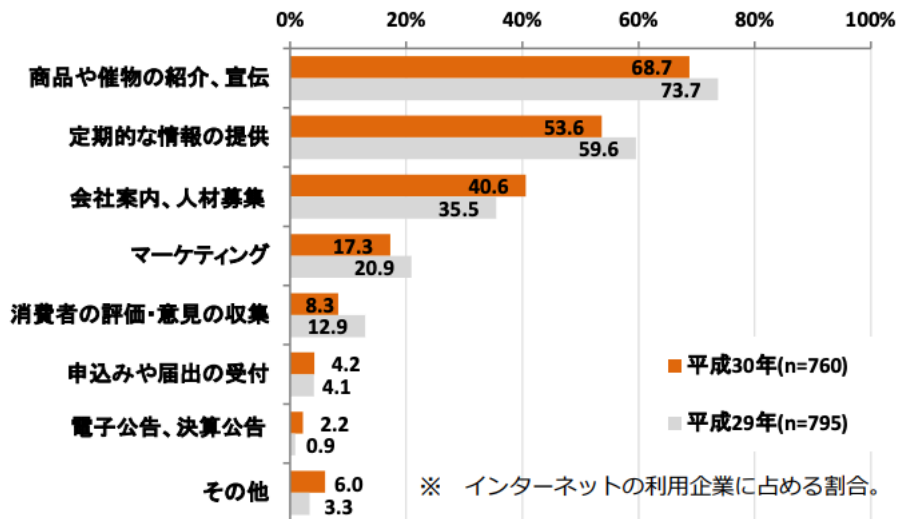


図 2.2 : SNS の利用目的 (企業) (平成 30 年, 平成 29 年)

2.2 用語

以下に、本研究で分析に用いた Twitter 用語の説明を示す。

- **ツイート**

Twitter に投稿するメッセージのことである。日本では「つぶやき」のことを指す。1 回の投稿で 140 文字 (英数半角で 280 文字) までの上限がある。テキストだけでなく、画像や動画も投稿が可能である。

- **タイムライン**

複数のツイートが時系列順で表示される場所のことを指す。Twitter にログインした際に表示されるホームタイムラインには、自身のツイートや、自分がフォローしているユーザー (フォロイー) のツイートの一覧が時系列順に表示される。ユーザーは他ユーザーをフォローすることで、このホームタイムライン上で、ツイートの閲覧ができるようになり、リツイートや、いいね、返信機能により他ユーザーとやりとりを行う。

- **フォロイー**

あるユーザーがフォローしているユーザーのことを指す。ここで、フォローとは、Twitter の機能の 1 つであり、自分以外他ユーザーのツイートを自分のタイムライン上に表示させる行為のことを指す。

- **フォロワー**

あるユーザーがフォローされているユーザーのことを指しフォロイーの対義語として使われる。例えば、A が B をフォローしているとき、A は B のフォロワーであるといえる。

- **非公開ユーザー**

ツイートをフォロワー以外に非公開に設定しているユーザー (アカウント) のことを指す。非公開ユーザーをフォローし、フォロー許可された場合のみ、ツイートの閲覧を行うことが可能となる。TwitterAPI では、データを取得したユーザーが非公開ユーザーであるかどうかの判別はつくが、フォロー許可されていない非公開ユーザーのツイートを取得することはできない。本稿では、非公開ユーザー以外のユーザーのことを公開ユーザーと呼称する。

- ユーザー名

英数字とアンダーバー (_) の計 15 文字以内で構成され, 「@○○」の形で表示される. その他アカウントと同一のユーザー名は登録することが不可能であり, ユーザーを特定する名前のような役割を持つ. ユーザー名は変更可能であり, 変更してもフォロー, フォロワー関係が変わることはない.

- いいね

Twitter の機能の 1 つで, 「お気に入り」とも呼称される. 自身のツイートや他ユーザーのツイートに対して, 共感したことや気に入ったという意志を示す際にユーザーが行う行為である. 2020 年 1 月現在の機能では, 自身のフォロワーが いいね を行ったツイートが「○○さんがいいねしました」とホームタイムライン上に表示されることが稀に起こる (図 2.3). 何を基準にツイートを選定し, これをユーザーのタイムラインに表示しているかのアルゴリズムは, Twitter 社側は公開していない. ユーザーは, ホームタイムライン上で他ユーザーのツイートを閲覧し, リツイートや, いいね, 返信などのアクションを起こすため, ツイートの拡散(リツイート)においてこの機能による影響も考えられる. 本稿では, この機能による拡散の影響についても言及する.



図 2.3 ホームタイムライン上のいいね表示

• トレンド

2020年1月現在、実装されている機能であり、今、何が話題となっているかをTwitter社が独自のアルゴリズムによって決定しランキング形式で表示したものである。ユーザーはこれによって、話題となっているトピックや、ツイートの閲覧が行える。日本のトレンドや世界のトレンドだけでなく、自分の住んでいる県や、トレンドが知りたい地域を設定すると、その地域で話題となっている情報の閲覧が可能となる。図2.4に千葉のトレンドを表示させた例を示す。トレンドはリアルタイムで変化し、他ユーザーが何に関心を持っているかが簡単にわかるため、人々の情報収集に役立っている。このトレンド経由でユーザーがフォロワー外の他ユーザーのツイートに対しリツイートやいいねなどのリアクションをとることがある。



図 2.4 トレンド表示 (Chiba のトレンドを表示させたとき)

• リプライ

Twitter上で他ユーザーのツイートに返信することを指す。返信先のユーザー名を含む@付きのツイートとして、表示される。ユーザーはリツイートやいいね、リプライによって、他ユーザーのツイートにリアクションを示す。本研究では、このリプライを除いたオリジナルツイートを分析対象とした。

• リツイート

Twitterの機能の1つであり、ツイートを再びツイートすることを指す。この機能によって、ツイートが自分とフォロワーのホームタイムラインに表示され、フォロワーとツイートを共有することができる。(図 2.5) ホームタイムライン上に表示されるフォロワーがリツイートしたツイートは、ツイート上部にフォロワーのなかのどのユーザーがリツイートしたかが表示されるため、どのユーザーが共有したいツイートか判別がつく。他ユーザーのツイートだけでなく、自分のアカウントのツイートもリツイート可能である。ただし、非公開ユーザーのオリジナルツイートはリツイートできないという特徴がある。また、公開ユーザーでフォロワーではないユーザーのツイートもリツイート可能である。リツイートは情報拡散において、非常に重要な役割をもち、フォロワーの少ない個人の発言が大きくリツイートを伸ばす事例もある。本研究では、このリツイート機能に着目し、分析を行い、その特徴を明らかにした。

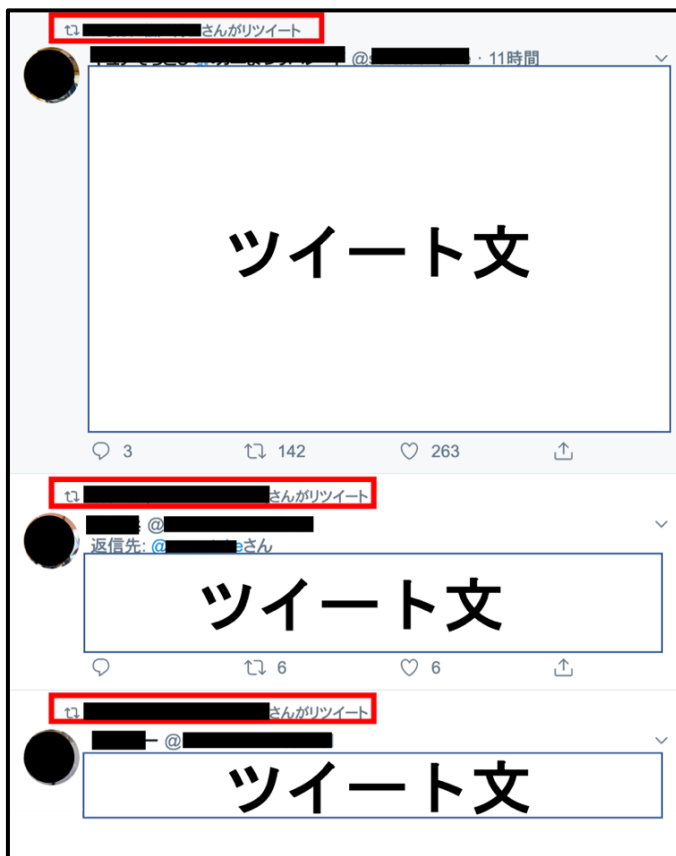


図 2.5 ホームタイムライン上のリツイート表示

2.3 TwitterAPI

2.3.1 概要

APIとは、「Application Programming Interface」の略語のことで、サービス提供者側の機能の一部を外部から利用できるように公開したもののことを表す。第3者開発したソフトウェアと機能を共有することを可能とする仕組みのことであり、ソフトウェアとソフトウェアを繋ぐ役割を持つ。GoogleやAmazon, Facebook, 楽天など多くの企業がAPIを活用している。TwitterもAPIを導入しており、Twitter社が提供しているAPIのことをTwitterAPIという。これは、Twitter社に利用申請を行い、アプリケーションの作成、APIを呼び出す際に用いる認証キーを発行してもらうことで利用可能となる。TwitterAPIは、多様なエンドポイントを含んでおり、ツイート情報の収集や、ツイートの投稿、返信、リイート、いいね、フォローなどの実行を公式のTwitter WEBサイトにアクセスせず実行できる。その他、ダイレクトメッセージ用のエンドポイントやTwitter内で広告キャンペーンの作成や管理を行う企業向けの広告用のエンドポイントが存在する。[8]このように、TwitterAPIは、企業、開発者、個人の利用者向けに提供されており、Twitter社への利用申請が通れば、誰でも利用することが可能である。使用の際は、目的に応じて必要なエンドポイントを利用する。本研究ではSearch APIというエンドポイントを利用し、データの収集を行った。

2.3.2 OAuth 認証

OAuthとは、ユーザーが情報を受け渡すときに行うアクセス権限の認可情報をやり取りするプロトコルのことである。セキュリティや利便性の面でOAuth認証が広く導入されている。Twitterでは、TwitterAPIの利用申請の際に作成・登録したアプリケーションへOAuth認証接続することで、TwitterAPIサービスを利用することが可能となる。データ収集・分析にはPythonを利用し、Pythonのoauth認証ライブラリ(requests_oauthlib)を活用してOAuth認証を行う。

2.3.3 パラメータとデータ構造

Twitter では、多くのオブジェクトを JSON で提供しており、TwitterAPI も同様に JSON 形式でエンコードされたデータを提供している。ここで、JSON とは、JavaScript Object Notation の略称であり、CSV や XML と同様のデータフォーマットの一種である。JSON のデータフォーマットは、key と value をペアで扱うという特徴がある。

TwitterAPI で取得可能なデータとしては、ツイートオブジェクト、ユーザーオブジェクト、エンティティオブジェクト、拡張エンティティオブジェクト、ジオオブジェクトがある。[9]ツイートオブジェクトとは、TwitterAPI でツイートデータを取り込む際に含まれるツイート情報のオブジェクトである。ユーザーオブジェクトは、ユーザー情報を含むオブジェクトのことである、ツイートオブジェクトにユーザーオブジェクトが含まれる。本研究では、このツイートオブジェクトのパラメータと、ユーザーオブジェクトのパラメータを利用して、データを取得した。以下に、ツイートオブジェクト内のツイートデータディクショナリと、ユーザーオブジェクト内のユーザーデータディクショナリの中で、データ収集の際に利用したパラメータとその説明を示す。また、図 2.6 に本研究で用いたパラメータをもとに、ツイートデータディクショナリの構造を表した図を示す。

ツイートデータディクショナリ

- id : ツイート ID (数値型)
- created_at : ツイートの投稿日時、UTC 時間 (協定世界時) は JST 時間 (日本標準時) に変換する
- text : ツイート本文
- ユーザーデータ辞書
- retweet_count : ツイートがリツイートされた数
- retweeted_status : ツイートがリツイートであるかどうかを判別させるパラメータ

ユーザーデータディクショナリ

- id : ユーザー ID (数値型)
- screen_name : ユーザー名
- followers_count : フォロワー数
- friends_count : フォロイー数
- status_count : ツイート数

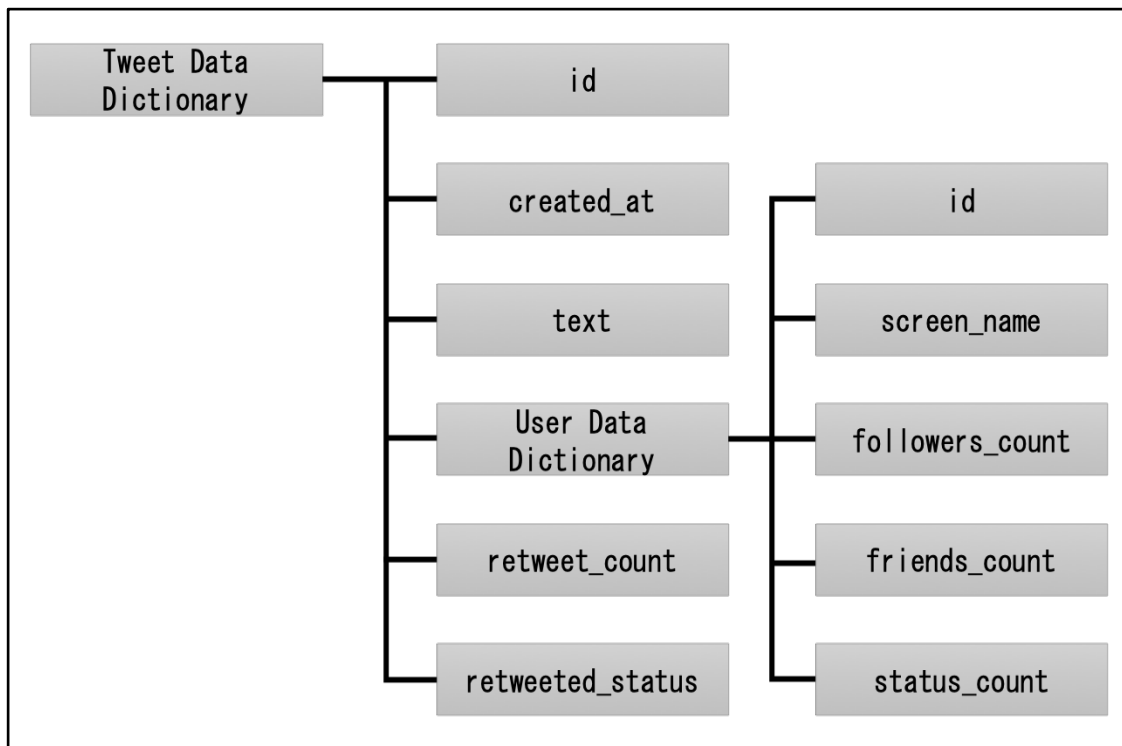


図 2.6 ツイートデータディクショナリの構造

TwitterAPI には、15 分毎に 180 回までというアクセス制限が設けられており、アクセス制限を超えてデータを取得することはできない。これをプログラム上で管理する必要がある。X-Rate-Limit-Remaining, X-Rate-Limit-Reset の 2 つのパラメータによって管理した。X-Rate-Limit-Remaining はアクセス可能回数、X-Rate-Limit-Reset はアクセスリセット時間を意味する。X-Rate-Limit-Remaining で残りのアクセス可能回数を取得し、これが 0 になると、API の使用制限に到達したことを表す。API の使用制限に到達するとアクセスが可能になるまで、スリープするように設定する。

第3章

リツイート特徴

3.1 フォロワー特徴分析

この節では、企業Aのアカウントを対象に、Aのフォロワーの中でAのツイートをリツイートした回数が多いユーザーの特徴を分析した結果を示す。この結果から、フォロワーの中で、どのようなユーザーがリツイートを多く行う傾向があるのかを考察する。

本研究で分析の対象としたツイートは、全てオリジナルツイートである。オリジナルツイートとは、リプライや他ユーザーのリツイートを含まないユーザー独自のツイートのことを指す。収集期間は、2019年9月13日から2020年1月12日までの4ヶ月間である。この期間のAのオリジナルツイート数は、466件である。以下に、分析対象としたAのフォロワー数やフォロワー数、4ヶ月間のツイートの平均リツイート数情報を示す。

フォロワー数:1727人 , フォロワー数 : 1777人 , 平均リツイート数 : 2.47件

以下に、分析の手順を示す。

1. TwitterAPIを用いて、4ヶ月間のオリジナルツイート全てのリツイートユーザーのユーザーIDを取得する。
2. 企業Aのフォロワー毎の情報(ユーザーID, フォロワー数, フォロワー数, ツイート数)を取得する。
3. 1と2のユーザーIDデータを比較し、リツイートユーザーのうち、Aのフォロワーだけを抽出し、フォロワー毎に、リツイート数, フォロワー数, フォロワー数, ツイート数の4つのデータをまとめる。
4. リツイート数とそれぞれの相関係数を取り、特徴を明らかにする。

ここでのリツイート数とは、ユーザーが4ヶ月間のAのツイートのうちリツイートしたオリジナルツイート数のことを指す。

図 3.1 に、ユーザーのリツイート数の補分布を示す。横軸にリツイート数 x 、縦軸に x を超える割合を対数目盛で示した分布グラフである。このグラフから、A のツイートに多く反応するユーザーはごく一部存在するが、フォロワーの中で、リツイートを行ったユーザーは少なく、ほとんどがリツイートを行っていないことがみてとれる。

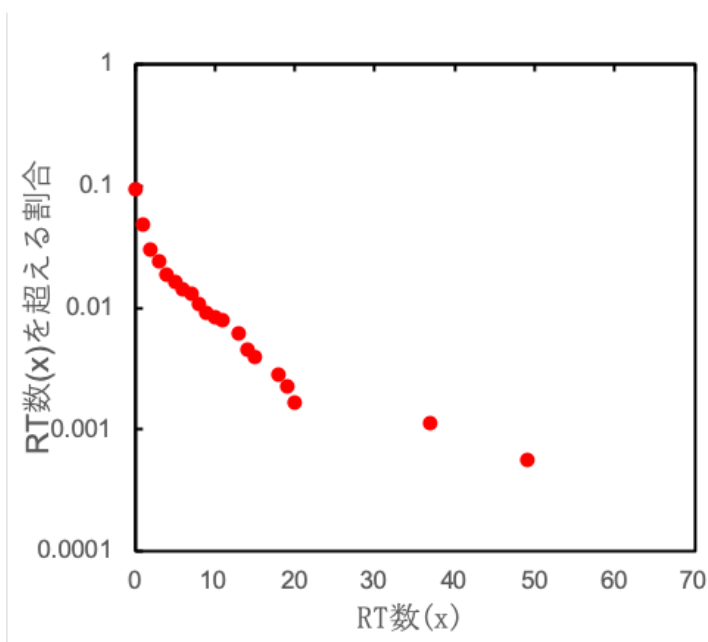


図 3.1 リツイート数の補分布

分析の手順 2 のフォロワー毎の情報は、2 章で示したユーザーデータディクショナリの `id`, `followers_count`, `friends_count`, `status_count` を用いて、取得した。また、表 3.1 に手順の 3 で得たフォロワー毎にまとめた結果をリツイート数の多いユーザー順に 20 件並べ、そのデータ例を示す。表の RT とは、リツイートの略語である。

取得したデータ、1777 人のフォロワーについて、リツイート数とフォロイー数、リツイート数とフォロワー数、リツイート数と総ツイート数の相関係数を算出すると、それぞれ、 -0.013 , -0.0068 , 0.18 となった。ほとんど相関関係はみられなかったが、フォロイー数との相関係数の方が、フォロワー数との相関係数と比較して、僅かに負の相関があることが示された。これは、フォロイー数が少ないユーザーほど A のツイートを見る機会が増え、リツイートを行う可能性が高いことを示唆している。また、ツイート数とは正の相関がみられ、ツイートを多く行うアクティブなユーザーほど、RT する傾向が高いといえる。

表 3.1 RT 数, フォロイー数, フォロワー数, ツイート数データ

ユーザー	RT 数	フォロイー数	フォロワー数	ツイート数
1	60	220	551	37959
2	49	490	283	32858
3	37	980	983	125813
4	20	1719	338	36405
5	19	628	560	3578
6	18	1617	480	57524
7	18	2097	1546	240397
8	15	1305	1065	329492
9	14	2367	277	50028
10	14	80	312	716
11	14	546	323	166566
12	13	773	503	39099
13	13	235	37	1956
14	13	4960	2951	153486
15	11	256	384	101230
16	10	1211	422	49505
17	9	5001	1576	96391
18	9	142	73	31297
19	9	1296	55070	56242
20	8	417	272	13393

3.2 フォロワー数とリツイート数関係

本研究では、前節で分析を行った A を含む 8 つのユーザーを対象に、2019 年 9 月 13 日～2020 年 1 月 12 日の 4 ヶ月のオリジナルツイートを対象にフォロワー数と平均リツイート数の関係を分析した。以下に、分析を行った 8 つのユーザーについての説明を行う。

分析ユーザー

- A～C：地域情報サイトを運営するアカウント
- D, E, G：地方自治体が運営するアカウント
- F：政治家が運営するアカウント
- G：球団関係者が運営するアカウント

これらのユーザーは、日々の情報発信や、サイトのアクセス数を増加させたいという目的で、アカウントを運営していると考えられる。そのため、ツイートの拡散力を高めることが求められている。

本節では、screen_name でツイートを取得するユーザーを指定し、2 章で示したツイートデータディクショナリの、id, created_at, text, retweet_count を用いて、各ユーザーのデータを取得した。図 3.2 にユーザー A の取得データの一部を示す。ツイート毎のリツイート数と、投稿日時データから、ユーザー毎に平均リツイート数を算出した。

tweet ID	created_at	text	RT
	2020-01-10 17:00:33		2
	2020-01-10 16:56:15		0
	2020-01-10 13:28:00		1
	2020-01-10 13:24:55		0
	2020-01-09 11:00:52		1
	2020-01-09 01:23:14		1
	2020-01-09 00:53:11		0
	2020-01-08 22:58:20		2
	2020-01-08 19:38:11		0
	2020-01-08 19:35:10		2
	2020-01-08 19:29:01		3

図 3.2 ユーザー A の取得データ例

表 3.2 にデータを取得した 8 つのユーザー毎のフォロワー数, 4 ヶ月間のオリジナルツイート数, 1 ツイートあたりの平均リツイート数を示す. 平均リツイート数の算出方法は, 以下に示した.

平均リツイート数 = 4 ヶ月間の合計リツイート数 / 4 ヶ月間のツイート数

表 3.2 ユーザー毎のフォロワー数, ツイート数, 平均リツイート数

ユーザー名	フォロワー数	ツイート数	平均リツイート数
A	1700	466	2
B	4300	1021	2
C	10400	361	20
D	47500	516	9
E	89400	246	151
F	164300	395	638
G	274300	250	342
H	765100	580	397

表 3.2 の結果をもとに, フォロワー数と平均リツイート数の相関係数を算出すると, 0.58 となり, 正の相関が見られた. 従って, 平均リツイート数の増加には, フォロワー数を増加させることが有効であるといえる.

第 4 章

拡散経路分析

4.1 分析手法

4.1.1 フォロイー・フォロワーネットワーク

この節では、拡散経路の分析のために取得したデータや、ネットワークの作成方法を説明する。ネットワーク作成は、以下の手順で行う。この手順により、作成したネットワークのことをフォロイー・フォロワーネットワークと呼称する。

1. 分析対象のオリジナルツイートのリツイートユーザー毎の、ツイート ID、ユーザー ID、リツイート時刻のデータを収集する。
2. リツイートユーザー全てのフォロイーリストを取得する。
3. リツイート時刻の早いユーザーから順に、ツイートのリツイートユーザーリストとリツイートユーザーのフォロイーリストを比較し、一致していれば、両者のユーザー ID を記録し、ネットワークリストに追加する。一致するユーザーが存在しなければ、情報源が不明であるとし、ネットワークリストに追加する。これを全てのリツイートユーザーに対して行う。

以下に、それぞれの手順をデータ例と共に詳細に説明する。

[手順 1]

手順 1 の実行結果の例を図 4.1 に示す。これは、TwitterAPI を用いて、データを取得し、csv 形式で出力したリツイートユーザーデータである。このデータ例をもとに説明を行う。「tweet id」のヘッダがある列には、tweet の ID 情報が書き込まれている。これは、ツイートデータディクショナリの id を指す。この列のヘッダを除く 1 番上の行は、オリジナルツイートのツイート ID が数値型で記載されている。リツイートであっても、ユーザー毎にそれぞれ異なるツイート ID が発行されている。2 列目は、retweeted_status によって、ツイートがリツイートかどうか示したものである。このパラメータは、ツイートがリツイートである場合に、ツイート元のオリジナルツイートの ID を返す。ツイートを検索して、リツイートユーザーを取得したデータであるため、2 列目の 2 行目以降は、オリジナルツイートのツイート ID が記されている。次

に、3列目は、created_at ツイート日時である。標準では、UTC(協定世界時)で出力されるため、JTC(日本標準時)に直し、出力した。リツイート時刻の早い順に上から並べている。4列目がツイート主とリツイートユーザーのユーザーIDである。このユーザーIDをもとに手順2でフォロワーリストを取得した。

tweet id	retweet id	created_at	user id
[REDACTED]65	[REDACTED]	0 2020-01-11 08:12:19	[REDACTED]
[REDACTED]61	[REDACTED]65	2020-01-11 08:12:44	[REDACTED]
[REDACTED]20	[REDACTED]65	2020-01-11 08:12:51	[REDACTED]
[REDACTED]52	[REDACTED]65	2020-01-11 08:12:53	[REDACTED]
[REDACTED]50	[REDACTED]65	2020-01-11 08:12:55	[REDACTED]
[REDACTED]96	[REDACTED]65	2020-01-11 08:12:56	[REDACTED]
[REDACTED]24	[REDACTED]65	2020-01-11 08:13:02	[REDACTED]
[REDACTED]48	[REDACTED]65	2020-01-11 08:13:07	[REDACTED]

図 4.1 リツイートユーザー情報取得データの例

[手順2]

手順1で取得したリツイートユーザーデータをもとに、全リツイートユーザーのフォロワーリスト(ユーザーIDリスト)を取得した。これは、図4.1の4列目の2行目以降のユーザーIDをもとに、フォロワーリストを取得したことを意味する。

[手順3]

ツイートのリツイートユーザーとリツイートユーザーのフォロワーリストを比較し、一致するとき、2人のユーザーIDをリストに追加した。一致しないとき、情報源不明として1を返した。その取得データ例を図4.2に示す。

2020-01-11 08:12:44	[REDACTED]	1	[REDACTED]
2020-01-11 08:12:51	114193	[REDACTED]	[REDACTED]
2020-01-11 08:12:53	114193	[REDACTED]	[REDACTED]
2020-01-11 08:12:55	114193	[REDACTED]	[REDACTED]
2020-01-11 08:12:56	114193	[REDACTED]	[REDACTED]
2020-01-11 08:13:02	114193	[REDACTED]	[REDACTED]
2020-01-11 08:13:07	114193	[REDACTED]	[REDACTED]

図 4.2 ネットワークリスト作成例

手順3のネットワークリストへの追加のイメージを図4.3に示す。ここで、ユーザーID:0000はオリジナルツイート投稿主のことを指す。図4.3はリツイート数3件のネットワークリスト追加のイメージ図である。このネットワーク図は、図4.4で示した。

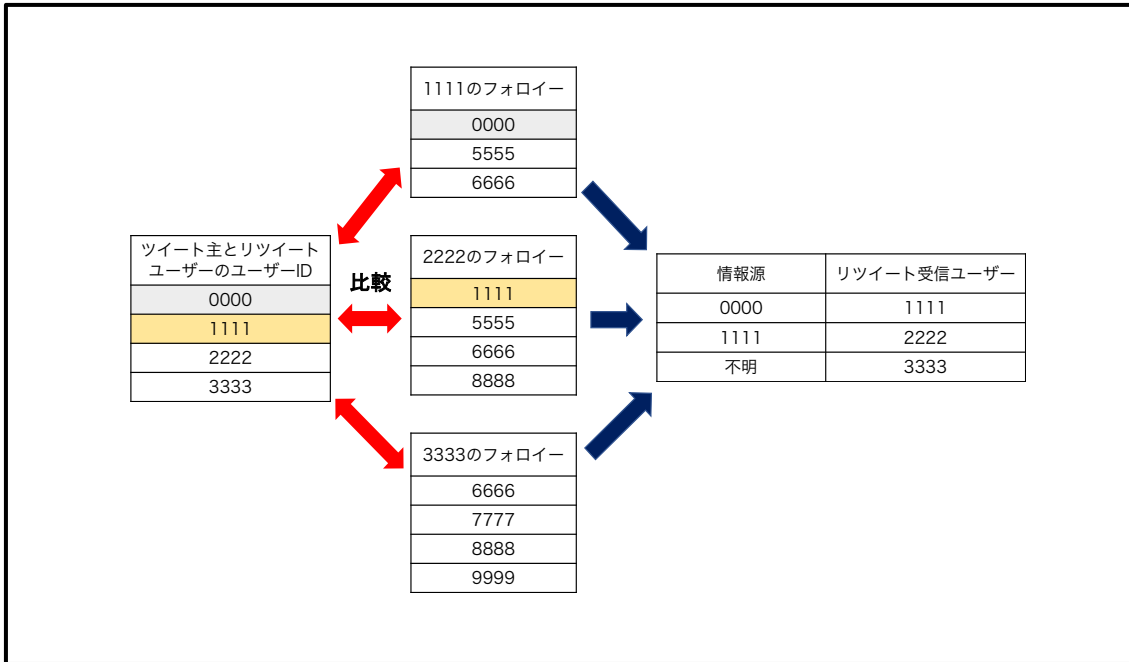


図 4.3 ネットワーク作成イメージ図

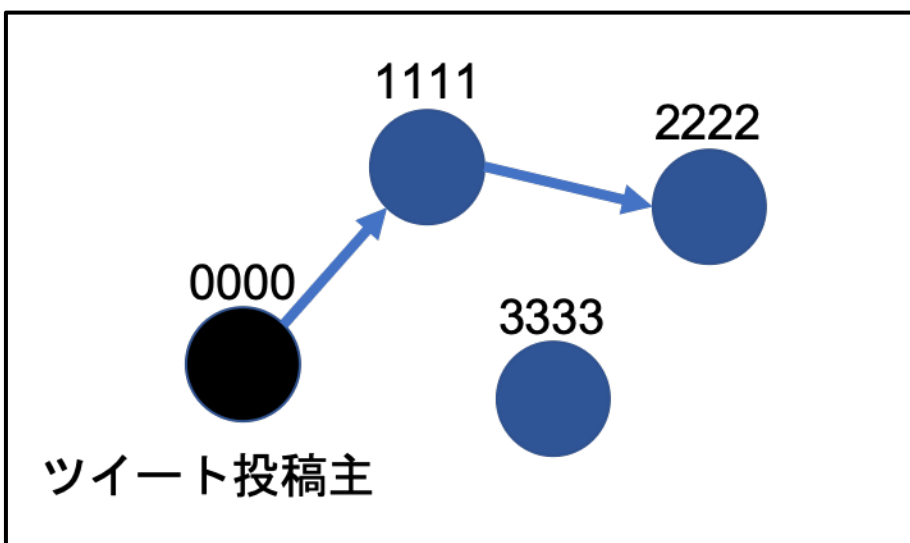


図 4.4 図 4.3 の作成ネットワーク図

4.1.2 1次拡散者, 情報源不明, 非公開リツイートユーザーの定義

本稿では, オリジナルツイートのリツイートユーザーのうち, フォロワーリストにツイート投稿主が含まれるユーザーのことを1次拡散者と定義した. 本研究では, ツイート毎にこの1次拡散者数, 情報源不明ユーザー数, 非公開リツイートユーザー数をだし, リツイート数との割合を算出することで特徴の分析を行った. この1次拡散者数は, 前項で示したネットワークリストの情報発信源がオリジナルツイートの投稿者であるリツイート受信ユーザーを数えることで取得した. また, 情報源不明ユーザーとは, 情報源が不明なリツイート受信ユーザーを指す. 情報源不明ユーザー数も同様に, 情報発信源が不明であるリツイート受信ユーザーを数え, 取得した. TwitterAPI では, ツイートをリツイートした非公開ユーザーのユーザーID やツイート (リツイート) ID を取得できない. そのため, 非公開リツイートユーザーの拡散経路の特定を行うことはできない. 非公開リツイートユーザー数については, ツイートのリツイート数から, TwitterAPI によって取得したデータ数 (公開リツイートユーザー数) を除くことで取得した. 本稿では, リツイートユーザーのうちの非公開ユーザーを非公開リツイートユーザー, 公開ユーザーを公開リツイートユーザーと呼称する.

図 4.5 に1次拡散者, 情報源不明ユーザー, 非公開ユーザーの説明をネットワーク図によって示した. この図は, リツイート数: 7, 1次拡散者数: 2, 情報源不明数: 1, 非公開ユーザー数: 1 というデータを示している. このネットワークリストは, 表 4.1 のように表される. 表には, 非公開ユーザーが含まれていないことがわかる. これは, TwitterAPI でリツイートユーザーを取得した際に, 非公開ユーザーが含まれないためである. 情報源不明ユーザーが非公開ユーザーからリツイートを受けとった可能性はあるが, 情報源が非公開ユーザーであるという特定はできない. 表 4.1 の不明以外のユーザー (ユーザー1, 2, 3, 4, 5, 6) は全て, 公開リツイートユーザーを指す.

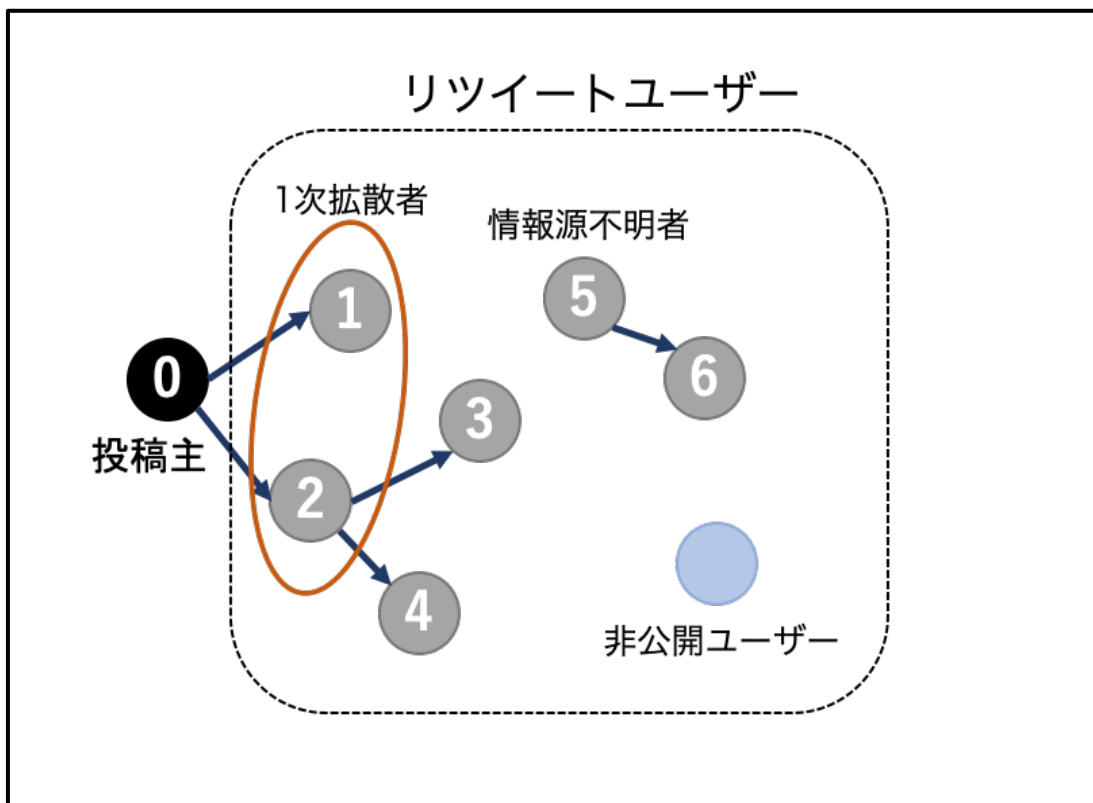


図 4.5 1次拡散者，情報源不明ユーザー，非公開ユーザー

表 4.1 ネットワークリスト

情報源	リツイート受信ユーザー
0	1
0	2
2	3
2	4
不明	5
5	6

情報源不明ユーザーの情報源として、本研究では以下を想定した。

- 「いいね」機能による表示
- 非公開ユーザーのリツイート
- Twitter トレンドや web サイト(ニュースサイト, まとめサイト)
- Twitter 検索機能

「いいね」機能による表示とは、ユーザーが他ユーザーからリツイート（ツイート）を受け取って、拡散したのではなく、「○○さんがいいねしました」とホームタイムライン上で表示されたツイートを見て、情報源不明ユーザーがリツイートを行ったことを想定している。

情報源が非公開ユーザーのリツイートについては、非公開ユーザーが TwitterAPI で取得できないため、情報源が不明となる。

Twitter のトレンドや web サイトとは、ユーザーが Twitter の機能であるトレンドに掲載されたツイートを閲覧しリツイートした場合や、Twitter 以外での web サイト(ニュースサイトやまとめサイト等)に掲載されたツイートをみて、リツイートした場合を想定している。これは、オリジナルツイート投稿主のフォロワー・フォロワーネットワーク経由外のユーザーが、ツイートを閲覧し、リツイートを行ったことを意味する。

Twitter 検索機能とは、ユーザー自身が興味あるトピックや、関心あるユーザーを検索し、ツイートを閲覧しリツイートを行った場合を想定している。ユーザー自身が興味あるトピックについて検索し閲覧するため、ツイートをフォロワーと共有する目的でリツイートを行う場合があることは考えられる。これについても、ツイートが投稿主のフォロワー・フォロワーネットワーク経由外で拡散されていることを意味する。

4.2 経路分析の結果

本節では、前節で説明した手順によって、いくつかのユーザーのツイートをもとにネットワークを作成した結果を示し、リツイート数増加の要因を考察する。

分析に用いたユーザーは、3.2節で示した8つのユーザーのうちのA, F, G, Hの4つである。Aについては、リツイート数上位20件のツイート、F, G, Hは、12月と1月中のツイートのリツイート数上位10件のツイートを経路分析の対象とした。

以下、取得したデータをもとに分析のため算出した割合を説明する。

- 非公開ユーザー割合 = 非公開リツイートユーザー数 / リツイート数
- 1次拡散者割合 = 1次拡散者数 / 公開リツイートユーザー数
- 情報源不明割合 = 情報源不明数 / 公開リツイートユーザー数

TwitterAPIの仕様上、非公開リツイートユーザーの経路は追跡不可能であるため、1次拡散者数や情報源不明数に非公開リツイートユーザーは含まれない。そのため、上記のように割合を算出し、分析に用いる。

4.2.1 リツイート数と各割合の関係

ユーザーAのリツイート数上位20件のツイートの経路を分析した結果と、リツイート数に対する非公開ユーザー割合、1次拡散者割合、情報源不明割合を算出した結果を表4.2に示す。また、F, G, Hについてもリツイート数上位10件のツイートの経路を取得し、それぞれの割合を算出した。F, G, Hそれぞれのユーザー毎に取得したデータ10件のうちのリツイート数上位5件を取り上げたデータの一部を表4.3に示す。表のRTとは、リツイートの略語である。また、図4.6にユーザーAのリツイート数と1次拡散者割合の散布図を示した例を示す。

表 4.2 ユーザーA の RT 数上位 20 件データ

ユーザー	RT 数 (x)	公開 RT ユーザー 数 (y)	非公開 RT ユ ーザー割合 ($[x-y]/x$)	1 次 拡散者数 (w)	1 次拡散 者割合 (w/y)	情報源 不明数 (z)	情報源 不明割合 (z/y)
A	42	39	0.07	16	0.41	3	0.08
	42	38	0.10	14	0.37	2	0.05
	25	21	0.16	14	0.67	2	0.10
	24	21	0.13	10	0.48	5	0.24
	20	20	0.00	11	0.55	1	0.05
	19	17	0.11	11	0.65	1	0.06
	19	15	0.21	7	0.47	6	0.40
	18	15	0.17	9	0.60	1	0.07
	18	16	0.11	10	0.63	2	0.13
	18	17	0.06	12	0.71	2	0.12
	17	15	0.12	7	0.47	1	0.07
	17	17	0.00	7	0.41	5	0.29
	16	13	0.19	9	0.69	2	0.15
	16	15	0.06	5	0.60	3	0.20
	14	11	0.21	4	0.45	1	0.09
	13	11	0.15	4	0.36	4	0.36
	13	12	0.08	6	0.33	2	0.17
	12	10	0.17	8	0.60	3	0.30
	12	12	0.00	5	0.67	1	0.08
	11	11	0.00	9	0.45	6	0.55

表 4.3 ユーザーF, G, H の RT 数上位 5 件データ

ユーザー	RT 数 (x)	公開 RT ユーザー 数 (y)	非公開 RT ユ ーザー割合 ($[x-y]/x$)	1 次 拡散者数 (w)	1 次拡散 者割合 (w/y)	情報源 不明数 (z)	情報源 不明割合 (z/y)
F	2372	1859	0.22	600	0.32	221	0.12
	752	645	0.14	379	0.59	146	0.23
	647	577	0.11	324	0.56	109	0.19
	423	364	0.14	222	0.61	67	0.18
	369	261	0.29	160	0.61	56	0.21
G	2026	1492	0.26	747	0.50	510	0.34
	816	610	0.25	405	0.66	101	0.17
	527	438	0.17	332	0.76	46	0.11
	439	324	0.26	195	0.60	45	0.15
	247	268	0.23	191	0.71	25	0.17
H	971	743	0.23	512	0.69	81	0.11
	470	364	0.23	315	0.87	28	0.08
	350	257	0.27	250	0.97	4	0.02
	308	255	0.17	226	0.89	22	0.09
	286	215	0.25	195	0.91	9	0.04

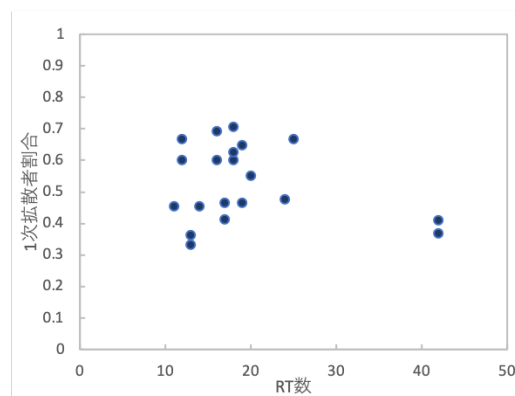


図 4.6 ユーザーA の RT 数, 1 次拡散者割合の散布図

ユーザーAについて

20件のデータから、リツイート数と1次拡散者割合の相関係数をとると、 -0.25 となり、負の相関がみられた。また、リツイート数と情報源不明割合の相関係数は、 -0.40 である。後述するユーザーF, G, Hの結果とは異なり、負の相関がみられた。これは、ユーザーF, G, Hのデータと異なり、ユーザーAのリツイート数上位のリツイート数が、最大でも42件と少なく、フォロワーネットワーク内でしか拡散が行われおらず、不明割合が低いことが考えられる。非公開リツイートユーザー割合は約0.10であることがわかる。

ユーザーF, G, Hについて

ユーザー毎に取得したデータ10件のリツイート数と1次拡散者割合の相関係数を算出するとそれぞれ、 -0.87 , -0.84 , -0.74 となり、ユーザーAの結果と同様に負の相関がみられた。また、リツイート数と情報源不明割合の相関係数はそれぞれ、 0.12 , 0.93 , 0.45 と正の相関がみられた。これは、負の相関がみられたユーザーAとは、異なる結果である。ユーザーFのツイート10件の1次拡散者割合をみると、リツイート数の最も多い2372件のデータは、1次拡散者割合が 0.32 と、他の9件よりも低い結果となった。これは、平均リツイートを大きく上回ってリツイートが伸びたツイートは、1次拡散者割合が低く、情報源不明割合が高くなる傾向にあることを示している。ユーザーHのツイート10件について分析しても、リツイート数が最も多い2026件の1次拡散者割合は 0.50 と、その他9件は 0.60 以上なのに対し、低い結果である。情報源不明割合をみると、最大リツイートのツイートは 0.34 と他の9件より高い傾向がある。ユーザーGについてもリツイート数の最大のツイートの1次拡散割合は 0.69 と他の9件よりも低く、情報源不明割合は 0.11 と高い傾向にあることが示された。また、非公開リツイートユーザー割合については、どのツイートにおいても $0.15 \sim 0.30$ 含まれることが明らかとなった。

次に、拡散経路を取得したユーザーA, F, G, Hのデータ全て(50件)についてリツイート数と、1次拡散者割合、情報源不明割合、非公開リツイートユーザー割合それぞれの関係を分析する。50件のツイートのリツイート数と1次拡散者割合の相関係数を算出すると、 -0.15 となった。また、リツイート数と情報源不明割合の相関係数は、 0.25 であり、正の相関がみられた。この結果から、リツイート数の多いツイートほど、1次拡散者割合は低くなり、それに伴い情報源不明割合が高くなる傾向があることがわかる。リツイート数の多いツイートほど、1次拡散者割合が低いとは、ツイート投稿主の直接のフォロワーが拡散している傾向が低いことを示している。情報源不明割合が高いとは、どこからリツイートを受け取ったか不明であるユーザーの割合がリツイート数の多いツイートほど高いことを示している。また、非公開リツイートユーザー割合については、どのリツイート数のツイートにおいても、 $15\sim 30\%$ 含まれることが判明した。これは、どのユーザーのツイートにもリツイート数によらず、一定数の非公開ユーザーが含まれることを示している。

従って、リツイート数を増加させるためには、この情報源不明の割合を増加させる必要があることが明らかである。情報源不明ユーザーの考えられる情報源は、前節で示したとおりである。このなかで、「いいね機能による表示」経由は、フォロワー・フォロワーネットワーク内で拡散が行われる特徴があること、「非公開ユーザーのリツイート」経由は非公開ユーザーのフォロワー数は一般に少ないため、リツイート数増加への影響が小さいことが考えられるため、これら以外の情報源であるフォロワー・フォロワーネットワーク外に拡散させることが、リツイート数増加に重要であると考えられる。オリジナルツイート投稿主のフォロワー・フォロワーネットワーク外のユーザーに拡散してもらうことが、リツイート数増加には重要であるといえる。

4.2.2 フォロワー数と各割合の関係

本項では、フォロワー数が少ないユーザーとフォロワー数の多いユーザーの拡散の違いを明らかにする。前項同様、ユーザーA, F, G, Hの50件のツイートデータの1次拡散者割合と情報源不明割合をもとに、フォロワー数との関係を分析する。表4.4にA, F, G, Hのそれぞれのユーザー毎のデータの一部、ユーザーAは10件、ユーザーF, G, Hは10件全てのデータを表4.4に示す。50件のデータのフォロワー数と1次拡散者割合、フォロワー数と情報源不明割合の相関係数を算出すると、それぞれ0.70, -0.40である。1次拡散者割合には正の相関が、情報源不明割合には負の相関がみられた。表4.4のユーザーHをみると、他のユーザーよりも1次拡散者割合が高くなっており、フォロワー数の少ないユーザーAと比較するとその違いは明らかである。ユーザーF, GについてもユーザーAと比較すると、1次拡散者割合が高い傾向にあることがわかる。この結果から、フォロワー数の多いユーザーのツイートは、1次拡散者が大半であり、情報源不明割合が低く、直接のフォロワーが主に拡散している傾向があることを示している。これは、フォロワー・フォロワーネットワーク内で2次、3次と拡散が広がる傾向が小さいといえる。フォロワー数の多いユーザーF, G, Hのツイートは、フォロワー数の少ないユーザーAよりもリツイート数自体は多いが、フォロワー・フォロワーネットワーク内でしか拡散が行われておらず、ネットワークが閉じていることを示している。これは、フォロワー数が多いからといって必ずしも拡散しているとは限らないことを示している。そのため、フォロワー数の多いユーザーについても同様にフォロワー・フォロワーネットワーク外に拡散させることが、リツイート数増加に重要であるといえる。従って、どのユーザーのツイートにおいても、投稿主のフォロワー・フォロワーネットワーク外のユーザーにリツイートしてもらうことが、効果的に拡散する上で重要であることを本研究で明らかにした。

表 4.4 フォロワー数と 1 次拡散者割合, 情報源不明割合

ユーザー	フォロワー数	1 次拡散者割合	情報源不明割合
A	1800	0.41	0.08
	1800	0.37	0.05
	1800	0.67	0.10
	1800	0.48	0.24
	1800	0.55	0.05
	1800	0.65	0.06
	1800	0.47	0.40
	1800	0.60	0.07
	1800	0.63	0.13
	1800	0.71	0.12
F	164300	0.32	0.12
	164300	0.59	0.23
	164300	0.56	0.19
	164300	0.61	0.18
	164300	0.61	0.21
	164300	0.80	0.11
	164300	0.90	0.09
	164300	0.91	0.09
	164300	0.85	0.08
	164300	0.83	0.13
G	274300	0.50	0.34
	274300	0.66	0.17
	274300	0.76	0.11
	274300	0.60	0.15
	274300	0.71	0.17
	274300	0.80	0.12
	274300	0.80	0.16
	274300	0.81	0.13
	274300	0.81	0.12
	274300	0.88	0.10
H	765100	0.69	0.11
	765100	0.87	0.08
	765100	0.97	0.02
	765100	0.89	0.09
	765100	0.91	0.04
	765100	0.86	0.08
	765100	0.97	0.02
	765100	0.89	0.08
	765100	0.99	0.01
	765100	0.85	0.09

第5章

結論

本研究では、いくつかの Twitter アカウントを対象にオリジナルツイートの分析を行い、拡散の特徴を明らかにした。

第3章では、企業Aのツイートを対象にフォロワー特徴分析（Aのフォロワーの中でAのツイートをリツイートする回数の多いユーザーの特徴分析）を行うことによって、フォロワー数の少ないユーザーや、アクティブなユーザーほど、リツイート回数の多い傾向にあることを示した。また、複数のユーザーを対象にフォロワー数と平均リツイート数関係を分析した結果、平均リツイート数増加には、フォロワー数を増やすことが有効であることを示した。

第4章では、複数のアカウントのツイートを対象に拡散経路分析を行い、以下を明らかにした。

- リツイート数の多いツイートほど、1次拡散者割合が低くなり、情報源不明割合が高くなる傾向にあること
- どのツイートにも非公開リツイートユーザーが一定数含まれること
- フォロワー数の多いユーザーのツイートが、必ずしも拡散しているとは限らないこと
- リツイート数増加には、フォロワー・フォロワーネットワーク外に拡散させることが重要であること

本研究では、フォロワー・フォロワーネットワーク外で拡散を行ったユーザーがどこからリツイートを受け取ったかの詳細は明らかになっていない。トレンドによる拡散の影響が強いのか、その他の情報源による影響が強いのかを明らかにすることが課題である。

また、今後は、さらに大規模なデータを用いたネットワーク分析や、数理モデルを用いてリツイート数増加法の検討、Twitter上の拡散現象のモデル化を行いたい。

謝辞

本研究を進めるにあたり，何度もご指導いただきました塩田先生に深く感謝申し上げます．また，研究室の方々にも，大変お世話になりました．ありがとうございました．

参考文献

[1]小出明弘, 斉藤和巳, 風間一洋, 鳥海不二夫, “情報拡散モデルを利用した超媒介者検出法,” 情報処理学会研究報告, Vol. 2013-ICS-170, No. 6, pp. 1-8, 2013.

[2]F. Cheong and C. Cheong, “Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods.” PACIS2011proceedings, 2011

[3]臼井翔平, 鳥海不二夫, “情報拡散に影響するネットワーク構造特徴,” 人工知能学会論文誌 Vol. 30, No. 1, pp. 195-203, 2015

[4]那須野薫, 松尾豊, “Twitter における候補者の情報拡散に着目した国政選挙当選者予測,” 人工知能学会全国大会論文集 Vol. 28, pp. 1-4, 2014

[5]南川雅人, 塩田茂雄, “ネットワーク上の情報拡散過程におけるノード相関の影響” 電子情報通信学会 コミュニケーションクオリティ研究会, CQ2017-58, pp. 43-58, 2017

[6]南川雅人, 中島圭佑, 塩田茂雄, “Twitter データの特徴分析と人間の行動モデル” 電子情報通信学会 ネットワークシステム研究会 NS-2018-290, pp. 557-562, 2019

[7]総務省 平成 30 年 通信利用動向調査の結果
https://www.soumu.go.jp/johotsusintokei/statistics/data/190531_1.pdf
2020/1/23 アクセス

[8]Twitter の API について
<https://help.twitter.com/ja/rules-and-policies/twitter-api>
2020/1/25 アクセス

[9]Tweet objects
<https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json>
2020/1/28 アクセス