Twitterにおけるデマユーザの抽出を目的とした特徴空間の構成

中川航至　播岡太朗　土方孝将　修兵
アドバイザー教員 遠藤靖典

1　はじめに

1.1　背景

Twitterとは，世界中の人々が参加するソーシャルネットワーキングサービスの1つであり，ユーザが「ツイート」と呼ばれる最大140文字の短文を投稿する ことによって，ユーザ間でそれらを共有することができるサービスである．サービス開始以来，携帯電話やスマートフォンなどのモバイル端末の普及によってユーザ数は増大し，2014年5月，全世界でのユーザ数は約2億2750万人に達した[1]．Twitterを使用することによって，情報をリアルタイムで多くのユーザに伝播する，あるいはそれを受け取ることができるため，ソーシャルメディアとしての注目も大きい．たとえば，2011年3月11日に発生した東日本大震災においては，連絡手段や情報入手ならびに情報発信の手段として注目を集めた．

Twitterは情報を高速で伝播させることができ，誰でもリアルタイムな情報を共有することができるという利点がある．一方で，伝播する情報の正当性は保証されておらず，特に緊急時において，デマ情報の拡散は不必要な不安を煽ることとなる．さらに，日本だけなく，世界で多くのユーザが利用していることから，デマ情報の拡散による風評被害や，インフラの混乱，それに伴う二次災害など，さまざまな場面に影響を及ぼすと考えられる．また昨今では，緊急時だけでなく，平常にもデマ情報の拡散による混乱がみられるケースがあり，Twitterを介したデマ情報に関して，ユーザが情報を正しく認知し，情報を発信，受信することが求められている．

1.2　目的

デマ情報のなかには，科学的・医学的知識不足により拡散されたものが存在する．たとえば,「ホウ酸を食べると放射線が防げる」というデマについて，ホウ酸による体内への影響に関する知識を有しているユーザは，この情報がデマであることを判断できる．そこで本稿では，ある情報に対して、その情報がデマであると妥当に判断を下すことのできるユーザを「健全ユーザ」と定義する．

現在Twitterは、一度発信したツイートを削除する機能を有する。これにより、健全ユーザがデマ情報をツイートしたユーザへ打消しツイートを送ることによって，デマツイートが削除される場合がしばしば生じるようになった．この機能の実装によって、Twitterはある種の自浄作用を有することになり、以前より信頼性のあるSNSになったということができよう。

しかしながら，打消しツイートを受け取ったにも関わらず、そのようなデマツイートを削除しないユーザも一定の割合で存在する．本稿では，そのようなユーザを「デマユーザ」と定義しよう．このようなデマユーザの発信したデマによって，Twitter上にデマツイートが残るだけでなく、時には拡散する．そこで、デマユーザに対しての対策として，例えばTwitter上の情報からデマユーザを判定し，デマ情報を拡散させないような伝播経路への介入が考えられる．

以上より，本稿では，ユーザ情報やそのユーザがツイートした文章からデマユーザの特徴を分析し，Twitterにおいてデマユーザの判定を行うことを目的とする．具体的には，Twitterで実際に投稿されたツイートやツイートを行ったユーザ情報などTwitter上から得られる情報からデマユーザの特徴を抽出し，評価・分析を行う．

1.3 Twitter の機能

Twitter には，ユーザ間のコミュニケーションを円滑に行うための，様々な特徴的な機能がある．その中から，本稿に関連する機能について，その概要を以下に述べる．

ツイート Twitter にメッセージを投稿すること，および投稿されたメッセージのこと．

リツイート 他人のツイートを引用して，自分のアカウントから発信すること．

フォロー 他のユーザのツイートを受信するように登録すること．また，フォローしているユーザをフォロワーという．

リプライ 特定のユーザ宛にメッセージを発信すること．相手のユーザ名の前に@をつけ,「@...」のように記述する．

また，自身で発信したツイートを Twitter 上より削除する機能を有する．例えばあるユーザのツイートした内容が他ユーザからの打消しツイートによりデマであることが分かれば，そのユーザが自身でそのツイートを削除することができる．

2 関連研究

本章では，ユーザがツイートした文章に関する分析やユーザ情報を用いた分析に関する文献について，その概要を述べる。

2.1 Twitter 上のデマ情報の検出

遠藤らは，文献 [2] で Twitter 上のツイートによって，デマを検出する手法を提案している．彼らは，Twitter 上からデマツイートを抽出し，そのツイートの分析を行った．ツイートの文章を形態素解析することによって分割した単語を肯定語，否定語に分類し，またツイートの拡散ネットワークに注目することによって，ツイートの伝播経路における情報を解析した．それら解析した情報を基にデマとニュースの分類指標を算出し，それらの指標を用いてノイズクラスタリングを行うことによってデマとニュースに分類した．また，ツイートのデマ度を推定することによって，デマとニュースを定量的に評価した．

2.2 自作自演ミームの検出に関する研究

Ratkiewicz らは文献 [3] で，ミームと呼ばれるネットワーク上を伝播するデマの検出に関する評価を行っている．この研究では，Twitter API から抽出したデータに対して監視，収集，処理を行う Thuthy システムの構築を行っている．このシステムの主な機能として，Web 上でユーザが収集されたデータに注釈をつける機能，自作自演ミームを検出する機能，拡散ネットワークの分析を行う機能を備えている．このシステムで取得したデータを分析することによって，自作自演ミームの拡散ネットワークは特異な特徴を持っていることを明らかにしており，明らかとなった特徴を利用した分類器を作成することによって，高精度で自作自演ミームの分類に成功している．

3 ツイート情報から得られる指標

3.1 Twitter API

本稿では Twitter API を用いてツイートデータを取得する．Twitter API は Twitter 社が提供するサービスで，ツイートデータやアカウント情報を取得することができる．Twitter API を用いることで，例えば特定のウェブサイトに特定のツイートが表示されるタイムラインを埋め込んだり，特定の内容を定期的に投稿するということが可能となる．

Twitter API にはアクセス過多によるサーバー負担を軽減するための取得数制限が存在する．これは取得するデータによって異なり，例えば特定のユーザのタイムラインからツイートを取得する場合，200 件のツイートを取得するプログラムを 15 分ごとに 15 回実行することができる．よって 15 分間に最大で 3000 ツイートを取得することが可能である．また，Twitter API は仕様の変更が行われることがあり，例えば自身のアカウントでリツイートしたツイートを取得することや，指定のユーザ同士のフォロー関係を直接取得する方法などがこれまでに廃止されており，実質的に取得不可能となった情報や間接的もしくは擬似的に取得しなければ得られないデータが存在する．特に現時点

(2014 年 10 月) ではリツイートに関する多くの関数が廃止され，リツイートのノードやエッジを直接的に取得することが難しくなった．それを踏まえた上で本稿では取得可能なデータを用いて特徴空間を構成し，デマユーザの特徴を調査する．

3.2 形態素解析

本稿では品詞の出現頻度を特徴の一つとして導入し，分析に用いるために形態素解析システムを利用している．形態素解析とは文章を単語ごとに分割し，品詞を判定することであり，ここでは MeCab[4] と呼ばれるソフトウェアを利用した．

3.3 データ収集方法

本稿ではニュースを発信するユーザ，健全ユーザ，デマを発信したことのあるユーザの 3 つの分類でツイートを収集し，分析を行った．デマユーザのツイートを抽出するために，本稿では Togetter[5] を用いてデマ情報を検索し，そのツイートを発信したユーザを特定している．該当したユーザのツイートを Twitter API を用いて抽出し，そのツイート群に形態素解析を行うことでツイートの文章の特徴を掴むための指標の元となる形態素を取得した．

3.4 ユーザ情報解析

ユーザの情報から, 特徴空間の軸の候補として以下を導入した．

表 1: 取得したユーザ情報

対象ユーザ数	39 ユーザ (メディア: 13, 健全: 13, デマ: 13)	
取得ツイート数	6900	
取得指標	フォロワー数/フォロー数	単語数
	極性 / 1 ツイート	肯定語 / 1 ツイート
	否定語 / 1 ツイート	肯定語+否定語 / 1 ツイート
	名詞数 / 1 ツイート	助詞数 / 1 ツイート
	動詞数 / 1 ツイート	形容詞数 / 1 ツイート
	接頭詞数 / 1 ツイート	記号数 / 1 ツイート
	肯定語+否定語 / 全単語	名詞数 / 全単語
	助詞数 / 全単語	動詞数 / 全単語
	形容詞数 / 全単語	接頭詞数 / 全単語
	記号数 / 全単語	

本稿ではこれら 19 の指標をもとに分析を行っている．

4 ユーザ特徴構造の分析

4.1 特徴指標の把握

ツイート情報から取得した 19 の指標をもとに，健全ユーザとデマユーザの特徴を把握する．ここで，メディアユーザについては，健全ユーザに似ているメディア，あるいはデマユーザに似ているメディアが存在していた．また，メディアユーザ，健全ユーザ，デマユーザについて，19 の指標をもとに 1 元配置分散分析を行ったところ，3 つのユーザの特徴が，メディアユーザもしくは，通常・健全ユーザの 2 つに分類されてしまう．これは，取得した 19 の指標の中で，明らかにメディアユーザが特質する指標が存在することが要因の 1 つと考えられる．たとえば，フォロー数に対するフォロワー数について，メディアは情報発信が専門であること，ユーザ側の視点からは，情報取得の重要なツールの 1 つであることから，フォローはしないが，フォローされるという特徴がある．本稿では，健全ユーザとデマユーザの特徴構造を把握することを目的としているため，これを把握することのできなくなる可能性があるメディアユーザは，分析から除外した．なお，メディアユーザについては，本章 3 節において，健全ユーザとデマユーザの構造の中で，どのような位置関係を取るのかということを理解する．

平均値の差異から特徴を把握するために，t 検定を行った．結果は表 2 の通りである．1% 水準で有意であったのは，名詞数，接頭詞数，5% 水準で有意であったのは，否定語数，肯定語+否定語数，記号数，形容詞率 (単語数あたり)，接頭詞率 (単語数あたり) であった．なお，形容詞率の値のみ，健全ユーザ > デマユーザ であったが，その他の値は，デマユーザ > 健全ユーザ であった．

4.2 健全ユーザとデマユーザの判別

前節で有意となった指標を用いて，従属変数をデマユーザであるかないか，とした判別分析を行った．このとき，指標の正規性は，Shapiro-Wilk の検定 (5% 水準) を用いて確認している．

まず，t 検定で有意となった 7 つの指標を用いて，強制投入法によって判別分析を行った．Wilks のラムダ検定により，このモデルが 5% 水準で有意でないこと

表 2: 指標間の t 検定結果

	t 値	df	平均値の差 (通常-危険)
フォロワー数/フォロー数	−0.17	24	−2.242
単語数	−1.95	24	−9.593
極性 / ツイート数	1.70	24	1.644
肯定語 / ツイート数	−1.18	24	−0.119
否定語 / ツイート数	−2.46	24	−3.403*
肯定語+否定語 / ツイート数	−2.43	24	−3.461*
名詞数 / ツイート数	−2.95	24	−1572.0**
助詞数 / ツイート数	−1.55	24	−373.692
動詞数 / ツイート数	−1.32	24	−146.615
形容詞数 / ツイート数	−0.18	24	−2.231
接頭詞数 / ツイート数	−3.14	24	−28.077**
記号数 / ツイート数	−2.42	24	−309.385*
肯定語+否定語 / 全単語数	−0.89	24	−0.143
名詞数 / 全単語数	−0.50	24	−0.013
助詞数 / 全単語数	0.68	24	0.010
動詞数 / 全単語数	1.38	19.021	0.011
形容詞数 / 全単語数	2.45	24	0.004*
接頭詞数 / 全単語数	−2.47	24	−0.001*
記号数 / 全単語数	−1.22	24	−0.010
*:$p < 0.05$, **:$p < 0.01$			

表 3: 判別分析結果

	正準判別関数係数	
	標準化された	標準化されていない
接頭詞数	0.682	0.03
否定語数	0.228	0.07
名詞数	0.177	0.13×10^{-3}
定数		−2.98
固有値 0.43, 正準相関 0.55,$\lambda = 0.70$, $\chi^2 = 8.10$, $p = 0.044$*		

が明らかとなった．よって，説明変数の見直しが必要となった．

この結果を踏まえて，ステップワイズ法による判別分析を試みた．しかしながら，1 変数のみが採択されるなどしたため，採択された変数を除外し，再度ステップワイズ法を用い，採択された変数を再度除外し，再々度ステップワイズ法を用い，という作業を繰り返し，最も影響をもたらすと考えられる指標を「否定語数」「接頭詞数」「名詞数」の 3 つに絞った．この 3 つの指標を説明変数として強制投入法で分析を行った．

Wilks のラムダ検定により，この 3 つを説明変数としたモデルが 5% 水準で有意であることが示された．結果は表 3 の通りである．標準化された正準判別関数係数より，判別への影響の大きさは，接頭詞数が最も大きく，次いで，否定語数となり，このモデルにおいては，名詞数が最も影響が低い．さらに，正準判別関数係数より，デマユーザ判別モデル y は次のようになる．

$$
\begin{aligned}
&y(\text{接頭詞数}, \text{否定語数}, \text{名詞数}) \\
&= \big(0.03 \times \text{接頭詞数} + 0.06 \times \text{否定語数} \\
&\quad + (0.14 \times 10^{-2}) \times \text{名詞数} - 3.002\big) \wedge 1 \vee 0
\end{aligned}
$$

$\wedge$ は論理積，$\vee$ は論理和を表す．また，接頭詞数，否定語数，名詞数は標準化されていない値を用いる。ここで得られた y の値はデマユーザの度合いを表し、0 に近いほど健全ユーザ、1 に近いほどデマユーザを意味する。

4.3 特徴構造の可視化とメディアユーザとの関係

前節の結果を踏まえて，特徴構造を可視化するために，判別分析に用いた 3 つの指標を組み合わせて散布図を作成した (図 1,2,3)．図の通り，これらの指標を軸としたグラフから，一部例外はあるものの，健全ユーザとデマユーザとの特徴傾向を把握することが可能であるといえる．今回の指標では，いずれも高くなればなるほど，デマユーザ傾向があることがわかる．

これらの傾向に加えて，13 のメディアユーザ情報を追加し，プロットを繋げた面として 3 次元で可視化したものが図 4 である．

ここからわかるように，今回用いた指標では，メディアユーザは一概に同様の傾向を取っておらず，健全ユーザに近いメディアもいれば，デマユーザに近いメディアもいることがわかる．また，健全ユーザならびにデマユーザのどちらにも近くないメディアユーザもいるということが明らかとなった．

5 考察

判別分析に用いた指標を見ると，否定語数，接頭詞数，名詞数である．否定語と接頭詞という点で見ると,“非” や “反”，“脱” といった単語が挙げられる．実際に，ツイート本文をみると，デマユーザのツイートには，“非武装”，“脱原発” などのワードが頻出し，さらに，名詞数が多いことを考えると，デマユーザは，1 ツイートに多くの情報を記しているといえる．これらを踏まえると，デマユーザは健全ユーザに比べ，自分の信条

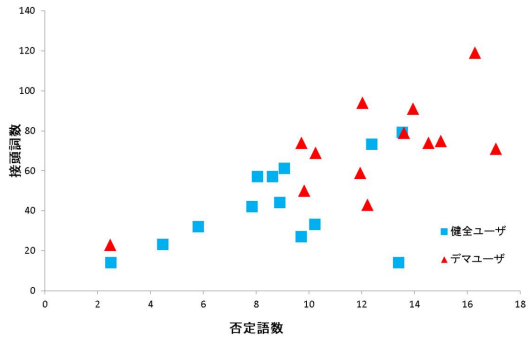

図 1: 否定語数 × 接頭詞

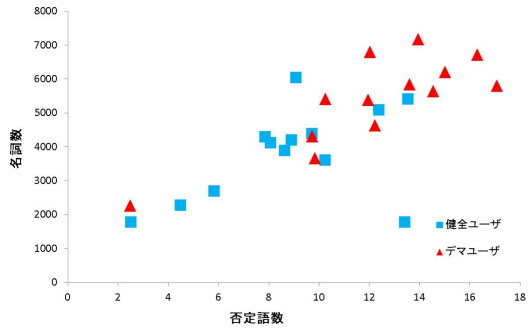

図 2: 否定語数 × 名詞

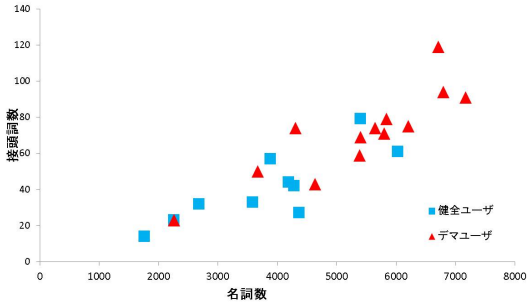

図 3: 名詞 × 接頭詞

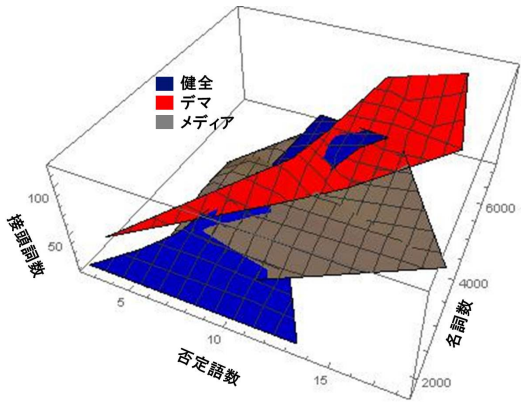

図 4: 3 次元空間へのプロット

や信念を強くツイートするために利用しているのではないかと推察できる．

また，メディアユーザについては，それ以外のユーザとは異なる特徴はあるものの，今回の分析で用いた指標で見ると，健全ユーザ，デマユーザどちらかに近づく傾向が示された．ここから，ツイート内容について，メディアごとに持っているスタイル（信条や信念）が要因の 1 つとなり，この結果を示したのではないかと考える．

6　おわりに

本稿では，健全ユーザ，デマユーザ，メディアユーザについて，ツイート情報を取得し，これらの情報から，特徴構造を把握することを目的として分析を行った．これにより，結果を次のようにまとめられる．

第 1 に，ツイート情報からユーザの特徴構造を把握できる可能性が示された．本稿では，フォロー数，フォロワー数，ツイート数というパーソナルな情報と，過去のツイートについて，形態素解析，極性辞書から 19 の指標を抽出した．これら 19 の指標のうち，7 つの指標では，健全ユーザ，デマユーザの傾向が異なることが示された．この要因の 1 つに，デマユーザと健全ユーザの間には，明らかにツイート使用，ツイート方法になんらかの違いがあることは容易に推察できる．

第 2 に，メディアユーザのツイートはメディア以外のユーザとその特徴構造が異なるということである．前述した通り，フォロワー数に対するフォロー数が明らかに異なり，これは，メディアユーザの使用目的が異なることが要因の一つに挙げられる．しかしながら，メディアユーザの中でも特徴があることも示された．判別分析に用いた 3 つの指標で見ると，健全ユーザに近いメディアユーザもいれば，デマユーザに近いメディアユーザもいる．この 3 つの指標は，ツイート本文の文章傾向に関するものであることから，メディアによって文章の特徴構造が異なるといえる．

今後の展望として，APIにより取得できるその他のツイート情報や，RT傾向，メンション傾向など，Twitter上のその他の影響を考慮する必要がある．本稿では詳細に調査をすることができなかったが，デマユーザは，信条や信念をツイートする傾向や，原発，憲法など，やや右寄りのツイートをする傾向があるのではないかと予測した．これらの要素を考慮し，健全ユーザとデマユーザの特徴について，さらに詳細に分類できるようにすることが可能になると推察する．

さらに，これらのユーザ特徴構造を用いて，ツイートのデマ度判定などにも発展させていくことで，デマユーザの判定をより信頼あるものとできるだろう．これらの試みは，Twitterの信頼性確保に大きく寄与すると考えられる．

謝辞

本演習を遂行するにあたり，工学システム学類4年の徳山晴紀氏には，ツイート情報の取得やTwitter API手法のご指導など，多大なご協力をいただきました．深謝いたします．

参考文献

[1] “Twitter ユーザは新興国で急伸，4年後にアジア太平洋地域が4割に”, 日経BP http://itpro.nikkeibp.co.jp/article/NEWS/20140528/559804/ (2014.6.2. 確認).

[2] 斉藤和孝, 田嶋脩平, 中村裕, 遠藤靖典, “Twitter上のデマ情報の検出”, 2011年度リスク工学グループ演習, 筑波大学大学院システム情報工学研究科リスク工学専攻 (2011).

[3] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Goncalves, Snehal Patil, Alessandro Flammini, Filippo Menczer, “Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams”, Proceedings of the 20th international conference companion on World wide web, pp.249–252 (2011).

[4] “MeCab”, 京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所, との共同研究ユニットプロジェクト.

[5] “Togetter”, http://togetter.com/