社会問題に関する情報カスケード検出

川本　貴史†　　豊田　正史††

† 東京大学大学院情報理工学系研究科　〒 113-0033 東京都文京区本郷 731 東京大学 情報理工学系研究科
†† 東京大学生産技術研究所　〒 153-8505 東京都目黒区駒場 461 東京大学 生産技術研究所
E-mail: †{kawamoto,toyoda}@tkl.iis.u-tokyo.ac.jp

あらまし　ソーシャルネットワーク上では単なる友人間の情報伝達にとどまらず，それが連鎖することによる大規模な情報カスケードが観測される．伝播される情報には様々な種類が存在し，マーケティングなどへの利用が考えられるが，社会的に影響力があるものは一部である．今回は Twitter から得られたリツイートについて伝播経路の構造を調査し，社会的に影響力のあるツイートを抽出する手法を提案する．

キーワード　Micro Blog，Twitter，ソーシャルグラフ, カスケード, 情報伝播, データマイニング

1. はじめに

友人関係をバーチャルに示すソーシャルネットワークが出現し大規模化するにつれ，単なる友人間での情報のやりとりにとどまらず，友人から受け取った情報をさらに自分の友人へと発信し，それが連鎖することによるソーシャルネットワーク上での大規模な情報拡散が観測されるようになった．このような情報拡散のことを情報カスケードという [1]．現在ソーシャルネットワーク上の情報カスケードを扱った情報伝達を定量的に研究が広く行われている [4]．情報伝達における研究には情報伝播の予測を行う研究や，情報伝播に対して強い影響力を持つユーザ (インフルエンサー) を特定する研究，ユーザの行動のモデル化を行う研究などが存在する．主にこれらの研究はバイラル・マーケティング，トレンド予測，個人に則した広告，スパム判定などへと応用が考えられている．しかし規模の大きなカスケードが全て社会的に影響力のある情報であるとは限らない．そのため，社会的に影響力のあるカスケードを抽出することは重要である．

そこで本稿では，現在最も規模の大きなソーシャルネットワークの一つである Twitter を用い社会的に影響力のあるカスケードの抽出を行うことを目的とする．Twitter が公式に提供する投稿を拡散する機能であるリツイートをカスケードとみなし，特に今回は

- 重要な報道に関するカスケード
- 報道に対する意見や批判に関するカスケード
- 注意喚起やデマ訂正を行うカスケード

を社会問題に関するカスケードとし，表 1 にそれぞれの例を示す．まず，一つ目のニュース自体を知らせるカスケードについてであるが，これは地震速報や，社会現象などの社会的に影響のある出来事を周知する例として非常に重要である．また二つ目のニュースに対する意見のカスケードであるが，これは報道に対する意見でユーザが特に共感するものであり，マスコミに対する批判であったり有識者による意見であったりする．これらは特に世論動向という意味で重要である．最後の注意喚起やデマ訂正を行うカスケードについてであるが，これにはスパムや災害に対する対処法を共有したり，誤った情報に対する訂正などが含まれる．こういった情報がユーザ間で広く共有された結果起こったカスケードであり，これらは社会的に与える悪影響を抑えるという意味で重要である．

表 1　社会問題に関するカスケードの例

カスケードの種類	内容
重要な報道に関するカスケード	【2013 年 1 月 14 日 13:44 JR 運行情報】 ▽運転見合わせ ▼中央・総武各駅停車 ▼総武快速線 ▼常磐線快速電車 ▼高崎線 ▼埼京線 ▼宇都宮線 .....
	遠地地震 1 月 5 日 17 時 58 分 ■震源■北米西部 ■マグニチュード■M 7.7 日本への津波の影響については気象庁が調べています。
報道に対する意見や批判に関するカスケード	沖縄の成人式。ニュースでは絶対に報道されないけど、新成人が国際通りのゴミ拾いをしています。暴れてる映像しかテレビでは映し出さないのね。 http://t.co/efVakP6c
	我が家にもテレ朝と日テレと TBS が取材申し込みにきて、政府と日揮は実名は公表しないと言っているのに、こんな卑怯なやり方で公表された情報に乗っかって、メディアとしてのプライドはないのかと言ってやりました。
注意喚起やデマ訂正を行うカスケード	【重要】昨日辺りから流行ってる「日頃の行いおみくじ」、どうやら勝手にフォローを行ったりツイートをする権限も解放してしまうスパムらしい。使ってしまった人は承認アプリリストから速やかに削除した方がいい。削除ページは公式サイトのココ→ https://t.co/mde5yDNy
	本日、18 歳以下の方が LINE を利用できなくなるというデマが出ていますが、そのような事実はありません。18 歳以下の方も引き続きご利用いただけますのでご安心下さい。公式情報はこのアカウントや公式ブログでご提供致します。 http://t.co/cZxJXXWf LINE

このように社会的に影響力のあるカスケードの抽出は重要であるが，この課題にはそもそも重要なカスケードは規模の大きなカスケードの中にわずか 1 割ほどしか含まれないという難しさと，140 字以内というツイートの文字制限によって本文からは特徴量がわずかにしか取ることができないという難しさが存在する．そこで本研究では，本文の情報に加えてカスケードのグラフ構造を特徴量に追加することで精度を向上させた分類手法を提案し，評価実験を行い，提案手法の有効性を示す．

2. 関連研究

マイクロブログのコンテンツを分類する研究は広く行われて

いる．Sriram らは，ツイートを 5 つの News, Opinions, Deals, Events, Private Messages の 5 つに分類するという手法を提案している [9]．この分類の際には分類の特徴量として Bag of words を用い，読みたい投稿を絞り込む際の手がかりとしてこの分類を用いることを想定している．しかし，これらの分類ではツイートの持つ情報の有用性について考えられておらず，また，分類の際にもグラフ構造の特徴量を用いていない．

また，Castillo らはツイートの信頼性を判定する分類器を作成している [3]．その際の分類器にはユーザの特徴量，トピックの特徴量，リツイートの特徴量を用い，J48Tree アルゴリズムにより決定木を作成している．また，その前段階としてそれらのツイートが News クラス，Chat クラス，判断できない，のどのクラスに属するかを判定する分類器も作成している．この分類器では本文の特徴量，ユーザの特徴量，トピックの特徴量，リツイートの特徴量を用いている．この研究においては，ツイートの信頼性判定ということに重点が置かれており，前段階の News クラス分類においても，特定の出来事に関するニュースかどうかということで判定が行われている．そのため，本研究の社会的に影響力があるかどうかという判断基準とは異なる．また，Castillo らが主に用いている特徴量はトピックやツイートの本文であり主にグラフ構造の特徴量を用いる本研究とは異なる．

一方で，カスケードのグラフ構造を用いた研究も多く存在する [6] [10]．その中で Cheng らはカスケードの成長予測にグラフ構造を用いている [5]．カスケードの成長予測の問題を定式化し，その上で時間に関する特徴量，ユーザに関する特徴量，構造に関する特徴量，コンテンツに関する特徴量を用い，分類器を作成した結果，時間，構造に関する特徴量が重要であるということを示している．本研究では社会的重要性についての分類の際にも，構造に関する特徴量を用いることで精度を向上させることができることを示す．

3. データセット

本節では本研究の実験に際して収集したデータについて述べる．分析にあたり，Twitter 社が提供する API を用いて，2012 年 1 月から 2013 年 1 月まで公開アカウントからツイートを収集した．その中でも Twitter が公式に提供しているユーザ間のインタラクションデータで，API によって取得できたもののみを用いた．さらにこのデータからインタラクショングラフとカスケードデータを作成した．

3.1 インタラクショングラフ

今回カスケードが伝播するネットワークとしてインタラクショングラフ G を作成した．インタラクショングラフとはユーザがノード，ユーザ間のインタラクションがエッジとなる有向グラフである．

3.1.1 Twitter におけるインタラクション

Twitter には，主にリプライとリツイートという 2 種類のユーザ間のインタラクションの方法が存在する．

- リプライ

@マークにユーザ ID を続けることでそのユーザに言及する投稿のことをリプライ (reply) という．リプライを受け取ったユーザは通常のタイムラインとは別にリプライに注目することができる．そのため主にユーザ間のコミュニケーションや，特定のツイートへの反応として用いられる．

- リツイート

他のユーザの投稿を自らの投稿として再投稿・拡散する機能のことをリツイート (retweet, RT) という．リツイートはユーザが興味をもった話題や意見を自身をフォローするユーザへ転送する目的や，投稿主に対する対話，意思表示として用いられる [7] [2].

3.1.2 インタラクショングラフの作成

カスケードの分析をするに際して，カスケードのグラフ構造を決定するためのユーザ間の関係を表すインタラクショングラフを作成した．このデータはカスケードデータとして用いる時期以前の 2012 年 1 月から 12 月のインタラクションデータを用いた．

今回作成したインタラクショングラフの統計量を表 2 に示す．今回分析の対象としたユーザは，この期間に一度はインタラクション元となっているユーザであり，およそ 100 万ユーザ存在した．また，インタラクショングラフとしてはメンション，リツイートのデータを共に用い，それぞれの有向エッジ数は 5000 万，1 億であり，実験の際にはリツイートに関してはリツイート元からリツイートしたユーザへ，メンションに関してはメンションされたユーザからメンションしたユーザへエッジが存在するとして，有向グラフを得る．これを解析に用いるグラフとして，そのグラフには 1 億 5000 万本のエッジが存在していた．

表 2　インタラクショングラフ

ユーザ数		1,066,870
エッジ数	メンション	58,627,341
	リツイート	114,848,093
	Mt & RT	153,711,945

4. 提 案 手 法

本章では，社会問題に関するカスケードを抽出する分類器を作成する際に用いた特徴量について説明する．まず実験対象としたカスケードについて説明した後，ベースラインとしての本文情報の特徴量，提案手法であるグラフ情報についての特徴量を述べる．その後，これらの特徴量を用いて作成した分類器について明し，提案手法の有用性を示すことを目的として行った実験の詳細を述べる．

4.1 カスケード

カスケードとして用いたのは，2013 年 1 月にツイートされ，多数リツイートされたツイートである．そのうち，前処理として，3.1 節で述べたメンショングラフに含まれるユーザのツイートに限定し，一回以上リツイートされたものは，17,508,576 種類であった．そのカスケードサイズの分布を示すグラフが図 1 であり，このうちの上位 500 種類のカスケードを以降の実験で用いた．これらのカスケードは表 3 に表すような本文のものが存在し，用いたカスケードの大きさは 8571〜1049 であった．

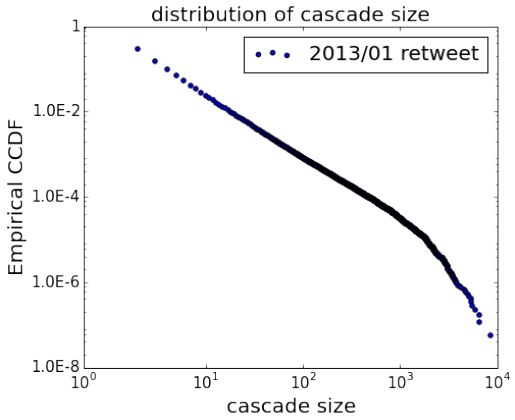

図 1　カスケードの大きさの分布

表 3　Top 5 cascade text

カスケードサイズ	本文
8571	髪の毛が後退しているのではない。 私が前進しているのである。 RT @kingfisher0423: 髪の毛の後退度がハゲしい。
6587	フクロウをいっぱい飼いすぎてしまうと、家に帰って来たとき、こんな出迎えを受けるらしいです…(´･ω･`)( ´･ω･`)(´･ω･`)( ´･ω･`) http://t.co/cMWRyBKP
6513	僕の猫、就寝時間を僕に合わせて行動するので、たまに夜更かしすると、待とうとしたのか、立ったまま寝ている事がしばしば。 http://t.co/CEOWBolu
5906	DOMDOM（ドムドムバーガー）を知ってる人いたら RT を。どれだけ知名度があるか気になる。
5430	最近読み始めた「ジョジョの奇妙な物語」の個性豊かな登場人物と、小鳥の可愛さをたくさんの人に知ってほしくて、思わず一つにまとめた。 http://t.co/o03QjYHC

表 4　社会問題に対するカスケードとそうでないカスケードの例

影響力のあるカスケード		
個数	分類	例
21	注意喚起やデマ訂正	こういう活動を皆さんに知っていただけたら嬉しいです。 http://t.co/8UooI0P
8	重要な報道	遠地地震 1 月 5 日 17 時 58 分 ■震源■北米西部 ■マグニチュード■M 7.7 日本への津波の影響については気象庁が調べています。
28	報道に対する意見や批判	1985 年の日本航空 123 便墜落事故で、事故からしばらくたった時、生存者の当時 12 歳の女の子がテレビの取材で「これから望むことは？」みたいなことを聞かれ、泣きながら「もうテレビが取材に来ないでほしい」と言ったのはいまだによく覚えています。
影響力のないカスケード		
個数	例	
443	「ゴキブリが彼氏ころしてくれた (o ´ ω ‘ o) 助かる」って呟き見つけて震えが止まらない (((˚Д ˚)))	
	パズドラプレゼント企画第 1 回目として、【魔法石約 80 個分×？人】･【魔法石 7 個分×？人】 を 抽選でプレゼントします !!! 今回のプレゼントは 1 回目なので参加者が多ければ当選者数も多くなります。 参加方法はＲＴ＆フォローだけ！ 【※参加受付・詳細発表は 7 日です】	
	DOMDOM（ドムドムバーガー）を知ってる人いたら RT を。どれだけ知名度があるか気になる。	

4.2　ラベル付与

4.1 節において作成したカスケードデータに対して，人手でラベル付けを行った．その結果の一部を例として表 4 に示す．ラベル付けは 1. で述べた通り社会的に影響力があるかどうかを基準に行い，URL が本文に含まれる場合は，そのリンク先も確認した上で行った．例えば表 4 の例の一番上の例であれば，このリンク先の画像は The yellow dog project という犬の保護活動を周知するポスターであるため，社会的に影響力があると考えられる．

ラベル付けした結果，社会的に影響力のあるカスケードの数は 57, そうでないカスケードの数は 443 となり，社会的に影響力のあるカスケードの数はそうでないカスケードの数と比較して，稀であるということがわかった．またこれ以降社会問題に関するカスケードのことを正解データ，そうでないカスケードのことを不正解データとする．

表 5　特　徴　量

本文情報に関する特徴量	
本文情報	各形態素が出現するか
URL	URL が含まれているか
グラフ情報に関する特徴量	
ルートユーザに関する特徴量	ユーザ ID
	$G(V_0)$ の Outdegree
	$G'(V_0)$ の degree
	$\hat{G}(V_0)$ の Outdegree
	$\hat{G}2(V_0)$ の Outdegree
リツイートしたユーザに関する特徴量	ユーザ ID の平均
	ユーザ ID の分散
	G の Outdegree の分布
	G' の degree の分布
	$\hat{G}$ の Outdegree の分布
	$\hat{G}_2$ の Outdegree の分布
	G' のクラスタリング係数の平均
	$\hat{G}$ のクラスタリング係数の平均
グラフ構造に関する特徴量	G' の最大の連結成分の大きさ
	G' の総エッジ数
	境界ノードの数
	境界エッジの数
	G' の深さの平均
	G' の深さの分布
時間情報に関する特徴量	ユーザがリツイートするまでの時間の平均
	ユーザがリツイートするまでの時間の分布
	ユーザがリツイートするまでの時間の変化
	カスケードの観測時間

4.3　本文情報に関する特徴量

表 5 に用いた特徴量の一覧を載せる．本文情報に関する特徴量としては，本文から URL を取り除き，Mecab を用いて形態素解析を行い，各本文にそれぞれの形態素が出現するかどうかを表す 1, 0 からなるベクトルを作成し，その上で URL が含まれるかどうかを示す次元を追加し，これを特徴ベクトルとして用いる．次元数は合わせて 4593 であった．

4.4　グラフ情報に関する特徴量

以下では 3.1 節で作成したグラフを $G = (V, E)$ と表現する．ここで，V はユーザの集合，E は 3.1.2 節におけるインタラクションの集合である．

4.4.1　カスケードに対して得られるグラフ

まずカスケード R_i はリツイートしたユーザ v と，そのリツイートした時間 t の組の集合として定義することができる．

$$R_i = \{(v, t) | v \in V\} \tag{1}$$

カスケード構造の例として図 2〜5 にノード v_0 のツイートが伝播するカスケードの例を示す．ノード v_i の添字 i は，リツイートした順番を示している．今回，一つのカスケードについ

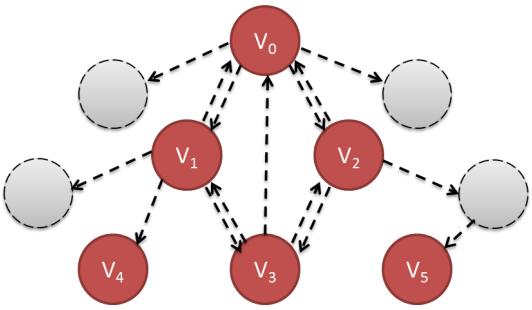

図 2　メンショングラフの例

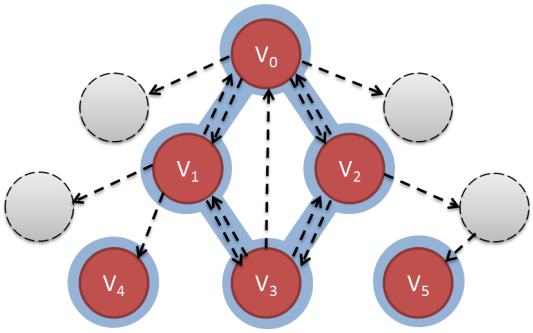

図 3　G' の 例

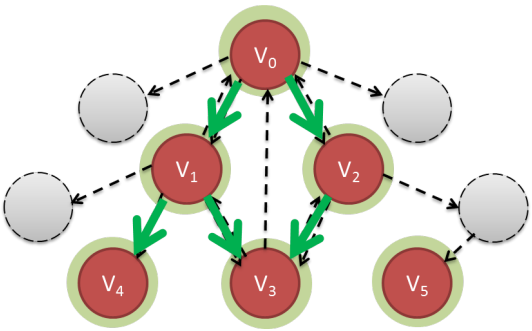

図 4　$\hat{G}$ の 例

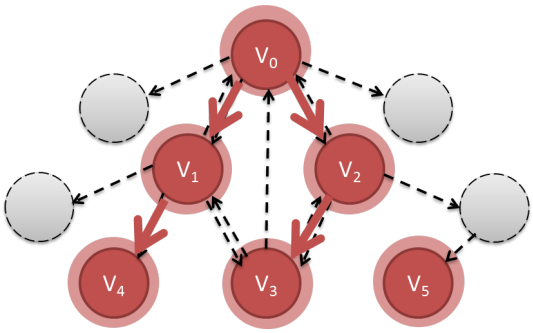

図 5　$\hat{G}2$ の 例

て，G 以外のグラフを複数考える．まず G' がそのカスケード R に含まれるノード v，それらのノード間のエッジのみからなる G のサブグラフのうち，双方向存在するノード間のエッジのみが存在するものとした無向グラフとして図 3 のように得られる．

$$G'(R) = < R_i, E'_i > \quad (2)$$
$$(\quad E'_i = \{(v_i, v_j) | v_i \to v_j \in E \cap v_j \to v_i \in E\})$$

次にカスケード R に含まれるノード v による G の部分グラフのうち，ノード間のエッジは添字の小さいノードから大きいノードへと有向であると考えたグラフとして $\hat{G}$ が図 4 のように得られる．

$$\hat{G}(R) = < R_i, \hat{E}_i > \quad (3)$$
$$(\quad \hat{E}_i = \{(v_i \to v_j) | t_i < t_j \cap v_i \to v_j \in E\})$$

さらに $\hat{G}2$ として，$\hat{G}$ のうち，全てのノードに対し入向辺を，支点の添字が最も大きいものだけが存在するとしたグラフが図 5 のように定められる．

$$\hat{G}_2\ (R) =< R_i, \hat{E}_{2i} > \quad (4)$$
$$(\quad \hat{E}_{2i} = \{(v_i \to v_j) | v_i \to v_j \in \hat{E}_i\} \cap \mathrm{Max}(v_i \in In(v_j)))$$
$$(\quad In(v_j) = \{v_i \to v_j | v_i \to v_j \in E\})$$

この時，G' はカスケードのノード間の順番を無視し，親密な関係のみを残したグラフ，$\hat{G}$ はカスケードのノード間の順番を考慮したグラフ，$\hat{G}2$ はカスケードは自分の直前にリツイートしたノードからのみ影響を受けるとして考えたグラフであるといえる．

4.4.2　特　徴　量

グラフ情報の特徴量として用いた特徴量を表 5 に示す．大きく分けて，ルートユーザに関する特徴量，リツイートしたユーザに関する特徴量，グラフ構造に関する特徴量，時間情報に関する特徴量を提案する．前者 2 つは，ユーザ自身の特徴とインタラクショングラフ上で直接接続しているユーザとの関係を表し，グラフ構造に関する特徴量，時間情報に関する特徴量はカスケードの伝播の特徴を捉えることを目的としている．

まずルートユーザに関する特徴量として，ユーザ ID，それぞれのインタラクショングラフの次数を用いている．ユーザ ID はそのユーザのアカウントが作成されてからの時間を表すと考えた．また，それぞれのインタラクショングラフの次数は，そのユーザの影響力の指標として考えられる．次にリツイートしたユーザに関する特徴量としては，まずユーザ ID の平均，分散を用いている．これらはアカウントの作成されてからの時期を表すため，ユーザの分布を示す指標として用いた．次にインタラクショングラフ上での次数の分布を用いた．この分布は次数の逆累積度数分布を次数の軸において対数軸で 12 個の bin に分け，特徴ベクトルの各次元に対応させるという手法で特徴量とした．次にグラフ構造に関する特徴量としては連結成分，総エッジ数についてはどれだけ密なグラフであるかという指標として，境界ノード，境界エッジは G において RT したユーザに接続しており，RT していないユーザ，またそのエッジの数として求めた．また，深さはどれだけルートユーザから遠くまで伝播したかという指標として用いた．時間情報についてはユーザがリツイートするまでの時間をルートユーザがツイートしてから経過した時間とし，それの平均，分布，変化を求め特徴量として用いた．分布については経過時間を時間軸で 8 個の bin に分け，それぞれの bin 毎のユーザ数を特徴ベクトルの各

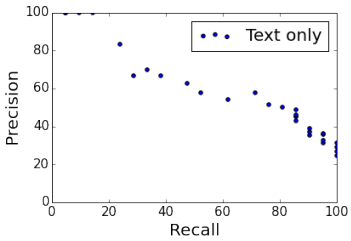

図 6　本文情報を用いた分類手法

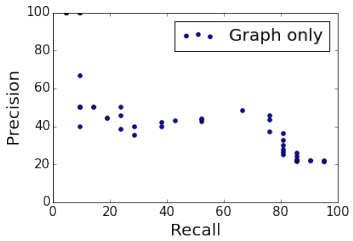

図 7　グラフ情報を用いた分類手法

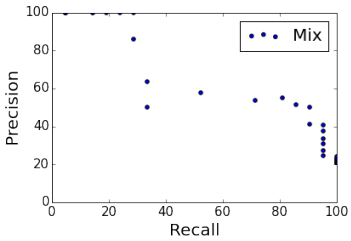

図 8　本文，グラフ情報を用いた分類手法

次元とした．

更に，カスケードの形状の時間変化を特徴とするため，グラフ特徴量をカスケードサイズが 5, 10, 20, 50, 100, 200, 500, 1000 の時点での特徴量を求め，これら全てを組み合わせたものをグラフ特徴量として用いた．

4.5　実験手順

まず，本文情報，グラフ情報それぞれの特徴量について比較を行うため，カスケードデータのそれぞれのカスケードの最初 1000 回のリツイートのみを取り出して実験を行った．なお，500 件のカスケードデータを時系列順に並べ，前の 400 件を訓練データとし，残りの 100 件をテストデータとした．まず訓練データから本文情報を用いた分類手法による特徴ベクトル，グラフ情報を用いた分類手法による特徴ベクトル，それらを組み合わせた特徴ベクトルそれぞれを用意し，RBF カーネルの SVM を用いて二値分類器を作成した．なお，本文情報を用いた分類手法による特徴ベクトルの要素は 0, 1 からなる一方，グラフ情報を用いた分類手法による特徴ベクトルの要素は実数値からなるためグラフ情報を用いた特徴ベクトルは正則化を行いモデルを作成した．続いてそのモデルを用いテストデータに対して分類を行った際の precision, recall,F 値について比較，評価を行った．

5.　実験による分類精度の評価

4.5 節に述べた手順で実験を行った結果を表 6 に示す．F 値による評価では，本文情報のみを用いた場合で F=55.0%, グラフ情報のみを用いた場合で F=58.2%，両方の特徴量を用いた場合で F=61.2%となった．この結果は，本文のみを用いた場合よりも Precision は劣るものの，Recall が大幅に増加し，またグラフのみを用いた場合よりも Recall は劣るものの，Precision が増加し，総合的には性能が向上したといえる結果となった．

また図 6〜8 に分離平面を変化させた時のそれぞれの Precision-Recall 曲線を示す．これらの図からはグラフ情報のみを用いた場合と比べると，右上に位置していることから性能が向上しているということができる．一方で本文情報のみを用いた場合と比べるとあまり変化は見られない．しかし，本文情報のみを用いた場合はトレーニングセット，テストセット間でのツイート内容の差に対応できておらず，最も良い分離平面を選ぶことができていなかった．その点，本文情報に加え，グラフ情報の特徴量を用いることで，ツイート内容に依存しないより一般的な分類器を作成することができたといえる．

次に両方の特徴量を用いて作られたモデルで，有効に働いた特徴量を本文，グラフそれぞれの上位 10 個を表**??**,**??**に挙げる．まず本文情報については句読点や「は」「です」など，公式の発表や，改まった文章に使われるような形態素が正例の判断に有効であり，負例の判断には「って」「たら」「？」「俺」などくだけた表現に使われるような形態素が有効であるというモデルとなっていた．一方でグラフ情報については $G', \hat{G}, \hat{G}2$ の Outdegree の分布については Outdegree が大きい bin の次元が正例の判断に有効で，小さい bin の次元が負例の判断に有効であり，G の Outdegree の分布については，グラフの大きさが 100RT までのグラフ特徴量に対しては最大の Outdegree の bin が正例の判断に有効であり，それ以降のグラフ特徴量に対しては負例の判断に有効となっていた．これは社会問題に対するカスケードは影響力の大きいユーザの RT を見て RT するユーザが多いからで，G の Outdegree がユーザの影響力を示すため，社会問題に対するカスケードは，比較的初期に影響力の強いユーザが RT することが多いと考えられる．他には正例への判断に有効である境界ノード，境界エッジの数は影響力の強いユーザがどれだけ RT しているかの指標として働き，社会問題に対するカスケードは影響力の強いユーザが多く RT していることを表している．

表 6　各手法の精度

	Precision	Recall	F score
Text	57.9	52.4	55.0
Graph	47.0	76.2	58.2
T & G	53.6	71.4	61.2

表 7　有効な本文特徴量

正例への判断	負例への判断
は	か
、	って
です	「
。	た
を	だ
よう	」
者	き
方	たら
【	？
】	俺

表 8　有効なグラフ特徴量

正例への判断	負例への判断
$\hat{G}2(V)$ の Outdegree の分布	$\hat{G}2(V)$ の Outdegree の分布
$\hat{G}(V)$ の Outdegree の分布	G' の深さの分布
$G(V)$ の Outdegree の分布	ユーザ ID の分散
$G\,'(V)$ の Outdegree の分布	$\hat{G}2(V)$ の Outdegree の分布
境界エッジの数	$G(V)$ の Outdegree の分布
$\hat{G}$ のクラスタリング係数の平均	$G\,'$ の最大の連結成分の大きさ
G' の深さの分布	G' の深さの平均
境界ノードの数	$G\,'(V)$ の Outdegree の分布
ユーザがリツイートするまでの時間の分布	ユーザがリツイートするまでの時間の分布
$G(V_0)$ の Outdegree	G' のクラスタリング係数の平均

6. まとめと今後の方針

今回，Twitter から得られたリツイートについて伝播経路の構造を調査し，社会的に影響力のあるツイートを抽出する手法を提案した．ツイートの本文の内容を用いた分類器としては形態素解析を行った単語の出現ベクトルを用い，伝播経路の構造として複数の伝播グラフを考え，それらに対して特徴を抽出することでグラフ構造の特徴を用いた分類器を提案した．結果としてはツイート内容のみを用いた分類器に比べて F 値の向上が見られ，自然言語情報だけでは分類が難しい問題も，グラフ構造を用いることで分類精度を向上させることができることが示された．

さらに，研究の発展としては精度の向上以外にも他の分野の研究を取り入れることも必要である．まずは伝播予測の研究であるが，今回調査した社会問題に対するカスケードと，そうでないカスケードを比較して，伝播予測に使用できる特徴量，伝播予測の精度に違いがでるかどうかを調査することが考えられる．他には，今回はリツイートによるカスケードを考えたためカスケードの内容の変化はなかった．しかし，投稿をクラスタリングした上で情報伝播の分析を行う手法が存在する [8] [11]．このような手法を追加で適用することで社会現象に対する反応も含めた調査が可能になる．それによって精度の向上や，更には時間変化を観測することでカスケードの内容の変化も扱うことが可能になることも考えられる．

文　献

[1] Bikhchandani, S., Welch, I. and Hirshleifer, D. A.: A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades ,1992.

[2] Boyd, D., Golder, S. and Lotan, G.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter, *2010 43rd Hawaii International Conference on System Sciences*, IEEE, pp. 1–10 ,2010.

[3] Castillo, C., Mendoza, M. and Poblete, B.: Information credibility on twitter, *Proceedings of the 20th international conference on World wide web - WWW '11*, New York, New York, USA, ACM Press, p. 675 ,2011.

[4] Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P. K.: Measuring User Influence in Twitter: The Million Follower Fallacy., *ICWSM*, Vol. 10, pp. 10–17 ,2010.

[5] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M. and Leskovec, J.: Can cascades be predicted?, *In Proc. WWW '14*, pp. 925–936 ,2014.

[6] Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z. and Kellerer, W.: Outtweeting the twitterers-predicting information cascades in microblogs, *In Proc. the 3rd conference on Online social networks*, USENIX Association, pp. 3–3 ,2010.

[7] Kwak, H., Lee, C., Park, H. and Moon, S.: What is Twitter, a social network or a news media?, *Proceedings of the 19th international conference on World wide web - WWW '10*, New York, New York, USA, ACM Press, p. 591 ,2010.

[8] Leskovec, J., Backstrom, L. and Kleinberg, J.: Meme-tracking and the dynamics of the news cycle, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, New York, New York, USA, ACM Press, p. 497 ,2009.

[9] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M.: Short text classification in twitter to improve information filtering, *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, New York, New York, USA, ACM Press, p. 841 ,2010.

[10] Weng, L., Menczer, F. and Ahn, Y.-Y.: Virality prediction and community structure in social networks., *Scientific reports*, Vol. 3, p. 2522 ,2013.

[11] 榊 剛史鳥海 不二夫 STAP 問題に対するソーシャルメディアにおける反応の分析，*WebDBForum2014* ,2014.