

推薦論文

マイクロブログからの社会的影響力を持つ 情報カスケードの検知手法

川本 貴史^{1,a)} 豊田 正史^{2,b)} 吉永 直樹^{3,c)}

受付日 2015年12月20日, 採録日 2016年4月13日

概要: マイクロブログではユーザ間での情報共有が連鎖することによる情報カスケードがしばしば観測され、その中には商品の回収につながりうる風評の拡散や災害への対処方法の共有など社会的影響力を持つものも存在する。このような情報カスケードに迅速に対応するために、本稿では教師あり学習に基づく分類器を用いて Twitter における情報カスケードから社会的影響力のある情報カスケードを検知するタスクを新たに提案し、これを機械学習に基づく分類器により解く手法を提案する。我々はまず、社会的影響力を有する情報カスケードの性質を理解するため、実際に拡散した社会的に影響力を持つ情報カスケードの内容の分析を行った。その知見をもとに、本文に含まれるテキスト情報や情報カスケードのグラフ構造の特徴量、情報カスケードごとのユーザ分布など、社会的影響力の有無の早期検知に有効であると考えられる多様な特徴量を設計した。実験では、実際のツイートデータセットにおいて検知する対象の情報カスケードのサイズ（初期リツイート数）を変化させて社会的影響力の有無を分類する実験を行い、早期検知の可能性を探るとともに、提案した特徴量の有効性を評価した。

キーワード: 情報カスケード, 社会的影響力, マイクロブログ

Detecting Information Cascades with Social Influence from Microblogs

TAKASHI KAWAMOTO^{1,a)} MASASHI TOYODA^{2,b)} NAOKI YOSHINAGA^{3,c)}

Received: December 20, 2015, Accepted: April 13, 2016

Abstract: In microblogs, information cascades caused by repeatedly sharing information among users often occur, and those information cascades sometimes have social influence in that they force companies to recall their products or they help user to know how to manage disasters. In order to quickly react to such information cascades, we set up a novel task of detecting information cascades with social influence from microblogs and use a supervised classifier to solve this task. We first investigate real cascaded information to understand what information has social influence. We then design features for a classifier using the textual content of the cascaded post, graphical characteristics of the developing cascade, and user distribution on the cascade that are effective for detecting social influence of the cascade. We apply our classifier to the information cascades in our Twitter dataset while varying the size of information cascade (early retweets) in order to evaluate the possibility of early detection of information cascades with social influence and the effectiveness of individual features.

Keywords: information cascade, social influence, microblog

¹ 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology,
The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
² 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo,
Meguro, Tokyo 153–8505, Japan
³ 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo,
Meguro, Tokyo 153–8505, Japan
本研究は著者が国立研究開発法人情報通信研究機構主任研究員を
兼務しているときに行われた。

1. はじめに

近年, Twitter や Facebook といったマイクロブログが

a) kawamoto@tkl.iis.u-tokyo.ac.jp

b) toyoda@tkl.iis.u-tokyo.ac.jp

c) ynaga@tkl.iis.u-tokyo.ac.jp

本稿の内容は 2015 年 11 月の WebDB フォーラム 2015 にて発表され, 同シンポジウムプログラム委員会により情報処理学会論文誌データベースへの掲載が推薦された論文である。

出現し、その上で友人関係をバーチャルに表すソーシャルネットワークが大規模化している。このソーシャルネットワークでは友人間でのコミュニケーションが行われるが、単なる友人間で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他の友人へと発信することが日常的に行われる。このような情報共有が連鎖することによって引き起こされる情報拡散を情報カスケードという [3]。

近年では、商品を批判する意見が拡散し、商品の回収につながるようなケースや、事故や災害への対処方法を共有する動きなど社会的影響力を持つ情報カスケードも観測されている。特にマイクロブログにおける情報カスケードには、マイクロブログのリアルタイム性の高さによって急速に広がる特徴があるため、政府、マスメディア、企業などにとって社会的影響力の高い情報カスケードを早期に発見することは風評被害に対する未然の対処、世論動向、報道、商品に対するフィードバックとして重要である。

以上のように、マイクロブログにおける情報カスケードには早期に対応、あるいは認知すべき情報が含まれるが、一方でアフェリエイトリンクへと誘導するスパムやジョーク、有名人の日常のつぶやき、広告など社会的影響力の少ない情報も多い。注目すべき情報カスケードを検知するための研究としては、将来的に広く拡散する可能性のあるカスケードを検知する手法 [7] や、スパムツイートを検知する手法 [6] があるが、前者は広く拡散する情報カスケードが必ずしも社会的影響力を持つとは限らない点で、また後者はスパム以外にも社会的影響力のない情報カスケードが存在するという点で、社会的影響力を持つ情報カスケードを検知する上では不十分である。

そこで本研究では、マイクロブログにおける情報カスケードの中から社会的影響力を持つものを検知するという新しいタスクを設定し、これを教師あり学習に基づく分類器を用いて具体的に解く手法を提案する。社会的影響力を持つ情報カスケードがどのようなものか自明ではないため、我々は「ツイートに含まれる情報が広まることで行動や意思決定に影響を受ける人が存在するか」という観点で社会的影響力を持つ情報カスケードの定義を行い、さらに被験者により社会的影響力を持つとされた情報カスケードを人手で分析することで、マイクロブログ上でどのような社会的影響力を持つ情報カスケードが存在するかを明らかにする。その後、この知見に基づき、社会的影響力を持つ情報カスケードの早期検知を試みる。3.3 節で明らかにするように、社会的影響力を持つ情報カスケードは全体の約 2 割程度しか含まれないという難しさと、投稿内容のみからでは分類のための手がかりが十分には得られないという難しさが存在する。そこで本稿ではテキスト特徴量に加えカスケードごとのユーザ分布やカスケードのグラフ構造を特徴量として用いることで社会的影響力を持つ情報カスケードの自動検知を試みる。

実験では、分析の際に構築したデータセットをもとに SVM を用いて分類器を学習し、分類器の示す F 値によって提案手法の有効性の評価を行った。

2. 関連研究

我々の知る限り、社会的影響力を持つ情報カスケードの検知を対象とした研究は存在しないが、情報カスケードの性質を明らかにすることを目的として、マイクロブログにおけるカスケードの予測に関する研究 [2], [7], [9], [16], マイクロブログにおけるカスケード分類の研究 [5], [11], [15], マイクロブログにおけるスパム検出などが行われている [6], [10]。以下で、これらのタスクについて、我々が提案するタスクとの関連性を明らかにするとともに、情報カスケードの性質の分析に用いられている特徴量を紹介する。

2.1 カスケード予測に関する研究

カスケードの将来の振舞いを予測する研究としては、カスケードの将来の大きさを予測する研究 [7] に限らず、実際につぶやくユーザを予測する研究 [9] や、ユーザの影響力を定量化する研究 [2] など幅広く存在する。また、対象とするカスケードもミームを Twitter のハッシュタグ [16] や URL [2], [9] とする研究や、あるいは Facebook の投稿拡散機能であるシェアによる情報カスケードを扱う研究 [7] など多岐にわたる。

カスケードの将来の大きさを予測する研究 [7] は、社会の中で広く拡散される投稿を当てる問題を解いており、ある意味では社会的に影響力の高い投稿を当てているともいえる。しかしながら、1 節で述べたように、広く拡散される投稿だからといって、必ずしも社会的影響力を持つ投稿とは限らない。一方で、我々の研究は、カスケードが成長するかどうかという問題自体は解いていないため、その点において、彼らの研究は我々の研究と相補的に用いるべき研究である。

2.2 カスケード分類に関する研究

マイクロブログにおけるカスケード分類の研究にはトピックによる意味的分類やグラフパターンによる構造的分類のほかにはツイートの信頼性判定を行うものなどがある。Sriram らはユーザが読むツイートを絞るための手助けとしてツイートを News, Opinions, Deals, Events, Private Messages の 5 つに分類する手法を提案している [15]。その際、分類の特徴量として Bag of words を用いている。また、Ren らは各ツイートに対し Web Forum や質問応答システムなどで一般的なラベル付けである複数の階層的な意味ラベル付けを行うことを提案している [13]。しかし、これらのタスクではツイートの持つ情報の社会的影響力については考慮されていない。

また、Castillo らはあるトピックのツイート集合におい

て 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した後, 自動でそのトピックの信頼性を判定している [5]. その際の特徴量としてはユーザーの特徴量, 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観点で判定が行われている. そのため, 本研究で考える, 情報が社会的に影響力を持つかどうかという判断基準とは異なる.

話題の早期検知を行う研究も存在する. このような研究は Topic Detection and Tracking と呼ばれ [1], ユーザごとのトピックや話題の時間変化をモデル化する研究や [4], 近年では SNS 上でのイベント検知を行う研究 [17] が行われている. これらの研究で対象とする話題やイベント自体の抽出は, 情報カスケード自体の抽出という点に関連するが, どの研究も情報の社会的影響力は考慮していない. 重要な話題の早期検知という観点では SNS 上で長期間流行する話題の早期検出を行う研究 [18], Twitter における政治的なトピックの発生を早期発見する研究 [14] など存在するが, これらと本研究とは注目する情報カスケードの性質が異なるといえる.

2.3 マイクロブログにおけるスパム検出に関する研究

近年マイクロブログでは悪意のある投稿やそれを自動で行うスパムユーザが増加しており, これらを自動検出することが広く求められている. Chen らはインタラクションの構造に着目し, クラスタリング係数や推移性などの指標が有効であるとしている [6]. Gao らは投稿に注目し, 投稿ごとにスパム判定をするシステムを提案している [10].

しかし, 1 節で述べたとおり, 社会的影響力の有無を判定するに際しては, カスケードがスパムでないと判定するだけでは不十分である. また, スパム検出は基本的にユーザや投稿に対して行っておりカスケードは対象としていない.

3. 情報カスケードへのアノテーション

本章では, 提案手法の評価に用いる社会的影響力の有無をアノテーションした情報カスケードのデータセットの構築方法とその内容の分析について述べる. アノテーション対象の情報カスケードとして, まず, Twitter API によるツイートの収集を行い, リツイート数に基づいて情報カスケードの抽出を行う. このようにして得られた情報カスケードについて, 被験者により社会的影響力の有無をアノテーションし, 評価用のデータセットを得る. その後, 社会的影響力を持つ情報カスケードの内容を分析する.

3.1 Twitter データセット

情報カスケードの分析, 評価を行うためのデータセット

表 1 情報カスケードの統計量

Table 1 The statistics of the information cascades.

	1 月	2 月
カスケード数 (600 リツイート以上)	1,130	475
カスケードへの参加総ユーザ数	137,167	70,922

としては, 著者らの研究室において 2011 年 3 月より継続的に収集している Twitter のデータセットを用いた. 本データセットは, 150 万人程度の公開ユーザからタイムラインを継続的に収集したもので, 2015 年 8 月時点で約 250 億のツイートが蓄積されている. 収集対象のユーザは, 2011 年 3 月に 30 名程度の著名な日本人ユーザを選択し, それらのユーザに対してメンション*1 やリツイート*2 を行ったユーザをさらに収集対象として順次拡大していったものである. この中から, 2013 年 1 月から 2013 年 2 月の間につぶやかれたツイートを用いて評価用データセットの構築を行った.

3.2 情報カスケードの抽出

本稿における情報カスケードは Twitter API による公式リツイートによって拡散されたツイート (元ツイート) とその公式リツイートの集合とする. 分析対象となる情報カスケードは, 次で述べるインタラクショングラフに含まれるユーザを観測対象のユーザセットと限定した上で 2013 年 1 月, 2 月のツイートそれぞれで 600 回以上リツイートが観測された日本語を含む元ツイートとそのリツイートを抽出することによって作成した. また, 非公式リツイートが起点となってリツイートされ拡散した情報カスケードは今回分析対象から外した. さらに, 情報カスケードの社会的影響力の有無を判断する際には (ツイート収集時に取得していなかった) リンクされている画像なども参照する必要があったため, リンク先を復元できなかったの 29 の情報カスケードは分析対象から外した. 今回対象とした情報カスケードの統計量は表 1 のとおりである.

インタラクショングラフの構築

情報カスケードを抽出する期間以前の 2012 年 1 月から 12 月のユーザ間の投稿のやり取りをもとにユーザ間の関係を表す有向グラフ (インタラクショングラフ) G を作成し, 情報カスケードの経路を推定する. なお, 推定した経路は次節で述べる提案手法で分類の手がかりの 1 つとして用いる.

リツイートとメンションはどちらもユーザ間の情報のやり取りを表しており, このようにして得られるユーザ間のつながりはカスケードの情報伝播の主要な経路となると考

*1 他のユーザに言及を行う機能であり, ツイートに対して意見や感想を述べるなどユーザ間のコミュニケーションに用いられる.

*2 他のユーザの投稿を自らの投稿として再投稿・拡散する機能であり, 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のに利用される.

表 2 インタラクシヨングラフの統計量

Table 2 The statistics of the interaction graph.

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

えられる。そこで各ユーザをノードとして（過去に）情報が流れる方向と同方向となるよう、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情報が流れるため同方向のエッジを、メンションに関してはメンションを送る際は送り先のユーザの投稿を見て送ったと考えられるためメンションの方向とは逆向きのエッジを追加しインタラクシヨングラフ G を得る。

今回作成したインタラクシヨ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ヨングラフ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。

3.3 被験者による社会的影響力の有無の注釈付け

3.2 節で得られた情報カスケードに対し、(著者を含まない) 3 人の被験者により、「ツイートに書かれた情報を知ったり、その情報を不特定多数に知られたりすることで、直接的あるいは間接的に行動や意思決定に影響を受けるか人がいるか」という観点で社会的影響力の有無を注釈付けした。注釈付けの際には、社会的影響力を持つ情報カスケードの典型例、およびボーダーケースとなる情報カスケードの例を含むアノテーションガイドラインを著者らが作成し、これを参照して実際にリツイートしたユーザが読む/見ると考えられる、元ツイートの本文、画像、元ツイートに含まれる URL のリンク先の情報を基に注釈付けを行ってもらった。これによって得られたラベルにおける 3 人の被験者間一致度 [8] は 0.69 となり、文献 [12] によれば相当な一致であり、社会的影響力を持つ情報に関してかなりの共通認識が得られていることが確認された。

最終的にラベルの不一致は多数決により解消した。その結果を表 3 に示す。表 3 から分かるとおり、社会的影響力を持つ情報カスケードは全体のおよそ 20%弱と少ないことが確認された。

3.4 社会的に影響力のある情報カスケードの分析

前節で抽出した社会的に影響力をもつ情報カスケードに対し、社会的影響力を持つカスケードにどのようなものが存在するかを明らかにするため、拡散された情報の内容について分類を行った。

分析の結果、社会的影響力のある情報カスケードは、1) 個人、組織、業界、社会への意見に対する共感・反感と、2) 影響力のある出来事（事実）の周知、として大別されること

表 3 1 月、2 月ごとのカスケードの社会的影響力の有無

Table 3 The distribution of the information cascades with social influence in January and February.

	1 月	2 月
社会的影響力有	188	106
社会的影響力無	942	369
合計	1130	475

表 4 意見の対象の事例

Table 4 Examples of the targets of opinions.

	明確	曖昧
個人	A 教授, B ちゃん, C 知事	渉外部 10 人のオヤジ, スポーツ記者, D 社社員
組織	E 社, F 新聞, G 番組	ある大手企業, テレビ, 学校, 朝の情報番組
業界	飲食業, マスコミ, ゲーム業界, 日本の IT 企業	
社会	高校大学生, 中高年, 韓国, 世論, 素人, 同性愛者	

が分かった。個人、組織、業界、社会への意見に対する共感・反感は、その意見に対し個人、組織、業界が対応する必要があるため社会的影響力を持ち、社会に対する意見では、そのカスケードが世論を反映していると考えられるため社会的影響力を持つ。一方、影響力のある出来事（事実）の周知には、事件・事故の速報や知られていない問題の周知が含まれ、それぞれ、世論動向や注意喚起、啓蒙やデマ訂正、問題提起として重要である。なお、この分類は第 1 著者が 1 人で行い、複数の分類にまたがると考えられる場合は最も顕著であると判断したカテゴリへと分類を行った。

なお、以後引用する事例では、特定の人名や企業名などが含まれるため、ツイートの内容が特定の個人・企業の誹謗・中傷につながると思われる場合は、著者の判断で <人名> や <企業名> のように匿名化を行った。

3.4.1 個人、組織、業界、社会への意見

個人、組織、業界、社会への意見に対する共感・反感に対しては、その意見の対象が何なのかを分類する。また対象が個人、組織のレベルである場合、対象の明確性によって情報カスケードに対する対応が変化すると考えられるため、対象が明確かどうかを分類した。表 4 にそれぞれの対象の事例を示す。以下でそれぞれの対象の粒度ごとに、元ツイートの例と社会的影響力があると判断される理由を分析する。

明確な個人 人名やスクリーンネームなど明確な個人に対して名指しで意見を述べている元ツイートが該当する。特に人数は 1 人とは限らず、グループで活動している場合はグループが対象にもなりうる。

きょうのシンポジウムの質疑応答で「自身の発言をどう考えているのか」という質問が<人名>に対して出て、匿名質問には答えないとつっぱねた。質問者は名乗り、それでも<人名>は答えず。きいていて思わず泣いてしまった。

この例では、本文に人名が出現しており、明確な個人の粒度であると判断される。そのうえで批判をされており、謝罪や釈明が必要になると判断されるため社会的影響力を持つ。

不明確な個人 意見の対象が個人であると判断できるものの、具体的な人物の特定まではできない元ツイート。

<人名>が授賞式に行くのにご本人の飛行機はエコノミー、ホテルは一泊 12000 円だったのに、必要もないのについてた英語 もしゃべれない渉外部 10 人のオヤジらの飛行機はビジネス、ホテルは一泊 5 万円だったんだと！(´Д｀) 言いふらしてやってくれと言われたので、言いふらすわ！

この例では、意見の対象は渉外部 10 人（不明確な個人）であるが、この行動に対して対応するのはこれらの人物が所属する組織であると考えられるため、その規模で悪評が広まるのが問題となる。

明確な組織 特定の企業や政府組織などが名指しで意見を述べられているもの。

< URL > 2 月 3 日の記事からご覧ください。全国チェーン、<社名>の対応です。私は許すことは出来ませんし、今後、関わることは一生ないでしょう。こんな会社が拡大することを望みませんし、全国のライダーにこの事実が届くことを切に願う。

この例では対象は<社名>であり、業務中の不手際を指摘されたものである。この情報が拡散し、社会的批判になると営業に支障がでる可能性があるため社会的影響力を持つ。

不明確な組織 明確な企業名などは明らかにされていないが、意見の対象が組織であるもの。

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。

この例では、意見の対象はテレビ局であり、具体的な局名は分からないため区分としては不明確な組織である。この意見はテレビ局全体の報道業界として対応す

る必要があると考えられるため社会的影響力を持つ。
業界 同じ産業、商売などに携わる組織、個人の集合が意見、批判の対象となっているもの。およびそこで共有される価値観や体制が意見、批判の対象となっているもの。業界全体として対処する必要のある問題である場合、社会的影響力を持つといえる。

<業界名 A>に転職したら人間として扱われるようになったのでカルチャーショックに驚いてる。仕事前に上司や先輩が怪我するなよって声かけてくる。やばい。熱でたら心配される。やばい。2 時間睡眠で動ける体になれて言われぬ。やばい。<業界名 B>勤務時代と全然違う。やばい。

この例では、本文で<業界名 B>の勤務体制が批判されているため、この区分に分類される。また、対象の業界全体として勤務体制の改善の必要があるため社会的影響力を持つ。

社会 国や年代などでくられた人の集団や、その間で共有されている価値観や体制が意見、批判の対象となっているもの。およびそこで共有される価値観や体制が意見、批判の対象となっているもの。

「ゲームでなら人殺しができる！」という人と、「ゲームの中でも人殺しはちょっと……」という人では、明らかに後者の方が「ゲームと現実の区別がついていない」のだが、世間的には何故か前者こそが「ゲームと現実の区別がつかなくなって殺人を犯す！」と嘯し立てられる。

この例では、ゲームによって犯罪が助長されるという風潮が批判されているため、社会の粒度に分類される。この問題は今後ゲーム規制などに発展する可能性があり、ゲーム業界にとって世論として重要である。

3.4.2 影響力のある出来事（事実）の周知

影響力のある出来事（事実）の周知に分類されるカスケードは、周知している出来事の事件性、その周知している出来事がどのような人、集団に対して重要なのかをもとに、速報、注意喚起・デマ訂正、啓蒙のいずれかに分類することができた。

速報 周知している出来事の事件性が高いもの。また、社会全体に対して大きな影響を及ぼすと考えられる出来事の周知。

【北朝鮮核実験情報】本日11時19分頃、気象庁が北朝鮮を震源とする地震波を観測、自然地震ではない可能性があります。北朝鮮による核実験の可能性もあるので、官邸対策室を設置しました。今後も随時情報をお知らせします。

この例は北朝鮮が核実験をした可能性があるという事件性の高い出来事を周知しており、重要である。

湘南新宿ライン、6日午前の運転取りやめ
<http://t.co/CEV7gUL9>

この例は翌日の鉄道の運行予測を周知するものだが、この運行予測によって影響が大きくなると思われるため社会的影響力を持つ。

注意喚起・デマ訂正 周知している出来事が特定のユーザーに対しての警告や対処法であったり、大きな影響を与える誤った情報の訂正である元ツイート。

【拡散希望】【警告！！】今年のインフルエンザは高熱が出ない場合を散見します。熱はそうでもないけれど喉が痛い・鼻水が止まらない等の風邪症状の方は、時間が許せば医師にインフルエンザの検査を受けてみてください。そして外出時のマスクと帰宅時食事前のうがい手洗いアルコール消毒をお願いします

この例は風邪症状の人々に対して、インフルエンザの検査をすすめるもので、インフルエンザの流行を抑え、正しい対処法を広めるということは重要であるため社会的影響力を持つ。

本日、18歳以下の方が<サービス名>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。<URL><ハッシュタグ>

この例は<サービス名>の利用制限が始まるというデマを訂正する元ツイートであり、18歳以下のユーザーが当該サービスを使い続けるかどうかに影響を持つ。

啓蒙 事件性のあまりない出来事、問題であるが、社会にあまり知られていない事実の周知。社会に対して有益な情報が拡散される場合や、問題の周知自体が目的のものが存在する。

表5 個人、組織、業界、社会への意見に対する共感・反感の分布
Table 5 The distribution of the targets of opinions towards persons, organizations, trades and society.

	明確	不明確
個人	23	33
組織	44	18
業界	32	
社会	45	

表6 影響力のある出来事（事実）の周知の分布
Table 6 The distribution of the notices of influential events (facts).

速報	36
注意喚起・デマ訂正	37
啓蒙	26

タバコの煙は、いま話題のPM2.5の一種に含まれ、喫茶店の喫煙席のPM2.5濃度は、中国の大気とほぼ同等。友人が参加した産業医の講習会で開会早々にあった衝撃的な話。

この例は対してタバコの煙がPM2.5の一種であるという社会にあまり知られていない事実を広める元ツイートであるため、啓蒙としての社会的影響力を持つ。

文章力を語る人は大抵まず語彙や知識に着目するけど、リズム感に関してはなかなか触れる人が少ない。語彙や知識が骨だとしたら肉となるのは間違いなくリズム感だと思う。リズムの生きた文章は滑るように読み進められる傍ら、リズムの死んだ文章は幾ら中身があっても疲れるだけで全く頭に入ってこない。

この例は文章力をつけるに当たって、リズム感という新しい観点からの評価の必要を訴えている。文章力の教育や、文章力をつけようと思っているユーザーにとって有用であり、社会的影響力を持つ。

3.3.2 項の分析を踏まえて著者が社会的影響力の種類を分類した結果、個人、組織、業界、社会への意見に対する共感・反感の個数の合計は195、影響力のある出来事（事実）の周知の個数の合計は99、社会的影響力のないカスケードの数の合計は1,311となった。社会的影響力を持つ情報カスケードの詳細な分類結果を表5、表6に示す。結果として、意見の項目では組織や社会に対する意見が多いことが分かった。特に世代などへの意見や、特定の企業に対する批判が多く見られた。また、被験者による社会的影響力のアノテーションの有無の不一致は啓蒙の項目に多く見られた。啓蒙では、社会的影響力の有無を判断する際に被験者自身の価値観に依存することになり、その部分のゆれが原因であると考えられる。

4. 提案手法

本研究では、RBF カーネルのサポートベクターマシン (SVM) を用いて情報カスケードの社会的影響力の有無を識別する分類器を学習し、これを社会的影響力を持つ情報カスケードの検知を行う。分類には、どのような内容の情報を、どのような過程で、誰が広めているかが手がかりとなると考えられる。そこで、これらをそれぞれカスケードした元ツイートから抽出したテキスト統計量、インタラクショングラフを利用して抽出したカスケードのグラフ特徴量、および拡散に参加したユーザーに関するユーザー特徴量でとらえることを考えた。以下で、順に説明する。

4.1 テキスト特徴量

テキスト特徴量としては (1) 出現する単語、(2) 固有表現が含まれるかどうか、(3) URL が含まれるかどうかの 3 種類を利用した。まず、本文に出現する単語の特徴量としては、Bag of words を用い、具体的には、元ツイートの本文を MeCab^{*3}で mecab-ipadic-NEologd^{*4}を辞書に用い形態素解析し、自立語の動詞、名詞^{*5}、形容詞を用いた。この特徴量により、事故や事件といった社会的影響力を持つ情報カスケードでよくみられる単語をとらえることができる。

3.4 節で分析したように個人や組織に対する意見では個人名や組織名などが固有表現として投稿に含まれる場合がある。そこで、固有表現抽出を行い、ツイート中の固有表現の有無を特徴量として用いた。固有表現抽出は CaboCha^{*6}を用いて行い、組織名 (ORGANIZATION)、人名 (PERSON)、地名 (LOCATION)、固有物名 (ARTIFACT) のいずれかの固有表現が含まれているかどうかで特徴量とした。

4.2 グラフ特徴量

カスケードの伝播経路構造を特徴量としてとらえることを目的にしてカスケードのグラフ構造を作成する。カスケードのグラフ構造はインタラクショングラフ G のリツイートしたユーザーの集合による部分グラフから作成されるが、親密なユーザー間でのやりとりに注目すること、仮想的な伝播経路を定めることを目的とし、エッジの残し方を変えることで 4 種類のグラフ構造を得る。まずエッジの残し方を、1) 以前にリツイートしたユーザーで、インタラクショングラフ上でエッジがあるユーザーすべてからのエッジを用いる場合、2) インタラクショングラフ上でエッジがあるユーザーのうち直前にリツイートしたユーザーからのエッジの

表 7 カスケード構造グラフ

Table 7 Cascade graphs.

	双方向エッジのみ	片方向エッジ
全ユーザ	$G'(R)$	$\hat{G}(R)$
直前ユーザ	$G'_2(R)$	$\hat{G}_2(R)$

表 8 グラフに関する特徴量

Table 8 Graph features.

(a) ルートユーザーに関する特徴量
Features on the root user.

$G(V_0), \hat{G}(V_0), \hat{G}_2(V_0)$ の outdegree
 $G'(V_0), G'_2(V_0)$ の degree

(b) リツイートしたユーザーに関する特徴量
Features on the retweeted users.

G, \hat{G}, \hat{G}_2 の outdegree の分布
 G', G'_2 の degree の分布
 G' のクラスタリング係数の平均

(c) グラフ構造に関する特徴量
Structural features on the cascade graph.

G' の最大の連結成分の大きさ
 G' の総エッジ数
 G' の深さの平均
 G' の深さの分布

みを用いる場合の 2 種類を用いることでインタラクショングラフ G の情報をそのまま残したグラフ \hat{G} と仮想的に伝播経路を定めたグラフ \hat{G}_2 を作成する。さらにそれぞれに対し双方向のエッジのみを残すことで得られるグラフ G', G'_2 を用いることでユーザー間の親密性をとらえたグラフを作成した。以上の 4 種類のグラフをまとめると表 7 のようになる。

今回用いたグラフ特徴量を表 8 に示す。大きく分けて、ルートユーザーに関する特徴量、リツイートしたユーザーに関する特徴量、グラフ構造に関する特徴量を提案する。前者 2 つは、ユーザー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ーとの関係を表し、グラフ構造に関する特徴量はカスケードの伝播の特徴をとらえることを目的としている。これにより、影響力のある出来事 (事実) の周知、具体的には速報や注意喚起・デマ訂正が、公式アカウントのような影響力の高いユーザーによって発信されることが多いという特徴をとらえることができる。

リツイートしたユーザーに関する特徴量としては、インタラクショングラフ上で度数の分布を用いた。この分布は度数の逆累積度数分布を度数の軸において対数軸で 10 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。また、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、深さはどれだけルートユー

*3 <http://taku910.github.io/mecab/>

*4 <https://github.com/neologd/mecab-ipadic-neologd/>

*5 ただし接尾辞、数は除く。

*6 <http://taku910.github.io/cabocha/>

ザから遠くまで伝播したかという指標として用いた。

4.3 ユーザ特徴量

マイクロブログではユーザごとにリツイートするツイートの内容に偏りがあり、社会的影響力のあるカスケードばかりをリツイートするユーザや、ネタなど社会的影響力のないツイートばかりをリツイートするユーザが投稿を拡散しているかどうか分類の手がかりとなる。そこで、テキストに対する Bag of Words を参考に、カスケードに参加するユーザを、いわば Bag of Users と見なして特徴量にした。

このユーザの特徴量の次元は、それぞれカスケードサイズ 600 の学習データに含まれるユーザに対応しており、次元数はカスケードサイズ 600 のもので 137,167 次元、50 のもので 33,234 次元となった。そして、あるカスケードのユーザ特徴量を作る際にはそのカスケードをつぶやいたユーザの対応する次元を 1、それ以外の次元を 0 にすることでユーザの特徴ベクトルとする。

5. 実験

本節では 2013 年 1 月のカスケードをもとに分類器を学習し、2 月のカスケードを自動で分類することによって、提案した特徴量でどれだけ正確に社会的影響力を持つ情報カスケードを検知することができるかを評価する実験について述べる。

5.1 実験手順

本研究の目的は早期の、つまりカスケードが広がらない段階でのカスケードの社会的影響力の有無の判断であるため、学習・評価の各カスケードから初期のリツイートを先頭から n 件を取り出し、これを初期カスケードと見なして分類器の学習・評価を行った (n=50, 200, 400, 600)。カスケードが広がるにつれ、グラフおよびユーザ特徴量については手がかりとしての有効性が増すと期待できる。

分類器の学習の際、グラフ特徴量は各次元が実数値を取るため、0 から 1 の間の値へ正規化を行い、分類器としては LIBSVM^{*7}を用いて RBF カーネルの SVM を学習した。また、分類のラベルに偏りがあるため、学習の際に正例側に重みをかけることで対応した。なお、パラメータチューニングは学習データにおいて 5 分割交差検定を用いて最大の F 値を取るパラメータを用いた。

本タスクには既存研究が存在しないため、どのような手法をベースラインとして設定するかが問題となるが、正例が少数であることからランダム分類のような通例よく用いられるベースラインは適切に機能しない。そこで、すべての情報カスケードが社会的影響力を持つとした場合を弱

表 9 カスケードサイズの変化にともなう F 値の変動

Table 9 F_1 scores while varying the cascade size

特徴量	カスケードサイズ			
	50	200	400	600
User	0.563	0.594	0.651	0.698
Text	0.552	0.552	0.552	0.552
Graph	0.468	0.521	0.515	0.503
All	0.576	0.614	0.661	0.709
Baseline	0.450	0.450	0.450	0.450
被験者 (A)	0.907	0.907	0.907	0.907
被験者 (B)	0.818	0.818	0.818	0.818
被験者 (C)	0.900	0.900	0.900	0.900

いベースライン (以下, Baseline) として採用し、提案手法により F 値でこれを上回ることを目標とする。加えて、4 章の各特徴量単独で分類を行った場合を強いベースライン (以下, Text, Graph, User) として、特徴量を組み合わせることの妥当性を確認する。

5.2 実験結果

特徴量を変えて実験を行った際の検知結果を表 9 に示す。参考のため、各被験者と正解ラベル (多数決) との一致度もともに示す。結果として、ユーザ特徴量とテキスト特徴量とグラフ特徴量を同時に用いたとき (All)、カスケードサイズがすべての時点で最も良い性能を示すことが分かる。また、これらの値はどれも、すべてのカスケードが社会的影響力を持つとした場合 (Baseline) の F 値 0.45 と比較して顕著な改善であることが分かる。また、カスケードサイズがどの時点でもユーザ特徴量はすべての特徴量を組み合わせた場合と比較しても遜色ない性能を示しているため、提案した特徴量のうちユーザ特徴量が性能において支配的であることが分かる。実用性を考えると、拡散初期の段階で社会的影響力を持つ情報カスケードを認識できることが望ましいが、カスケードサイズを減らし、拡散初期に利用できる特徴量を用いた場合でも、ユーザ特徴量に加えてテキスト特徴量やグラフ特徴量を用いることで、0.576 と Baseline を上回る F 値で社会的影響力の有無を判定できている。

また提案手法は、最も F 値の高いカスケードサイズ 600 のときで 0.709 であり、最も正解と一致しなかった被験者の F 値 0.818 に対してはまだかなりの開きがある。提案手法では被験者が参照していないユーザやグラフに関する特徴量を参照している点において被験者より有利であるが、URL のリンク先 (特に画像) を情報として利用できていないため、この点で差が出たと考えられる。5.3.2 項でこの点について考察を行う。

*7 <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

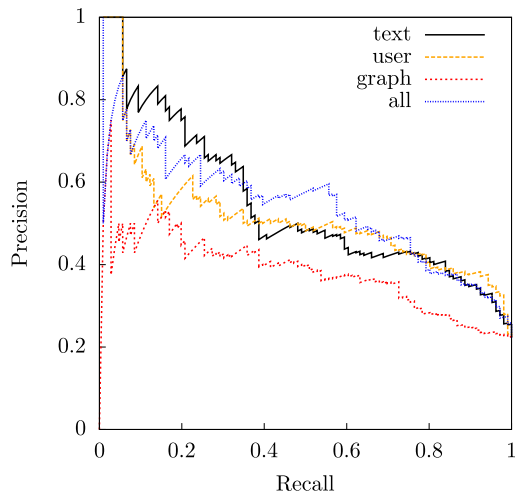


図 1 先頭 50 ユーザによる分類結果

Fig. 1 Classification results of information cascades using the first 50 users.

5.3 考察

5.3.1 カスケードの早期検知に貢献する特徴量の分析

次に、各カスケードを SVM で分類する際の判断基準である、分離超平面からのマージンの閾値を 0 から動かすことで、分類器の精度と再現率のトレードオフを調査した。図 1 にカスケードサイズが 50 のときの適合率と再現率のトレードオフを示す。テキスト特徴量は Recall が 0.4 以下と低い段階で、ユーザ特徴量と比べて良い性能を示している一方で、Recall が 0.4 程度の時点で性能に大幅な低下が見られる。ユーザ特徴量はテキスト特徴量と比べて Recall が上がった状態での性能の低下が少なく、テキスト特徴量と併用することで分類器の性能が向上したと考えられる。

5.3.2 URL を含まないカスケードについての分析

テキスト特徴量の性能が Recall が 0.4 程度の時点で性能の大幅な低下が見られる原因を探るため、カスケード本文の URL の有無について分析を行った。2 月のデータで本文に URL が含まれるカスケードは全体 475 件中 252 件と全体の半数以上であり、その多くは画像を引用したものであった。カスケードに添付された画像が投稿の社会的影響力の有無を判断するのに必要となる場合、テキスト特徴量は役に立たないと考えられる。そこで、URL を含まない情報カスケードについて提案手法の性能を分析する。

図 2 に URL を含まないカスケードについてカスケードサイズが 50 のときの適合率と再現率のトレードオフを示す。図から分かるとおり、図 1 と比較してテキストのみを手がかりに用いた場合の分類性能が大きく改善していることが分かる。つまり、テキスト特徴量のみを用いた分類器の分類性能低下の原因は、リンクされた情報(特に画像)によって社会的影響力を持つカスケードが存在するためだといえる。そのため、今後性能を改善させるためにはツイートに含まれるリンク先の画像の特徴量など新しい特徴量を設計する必要がある。

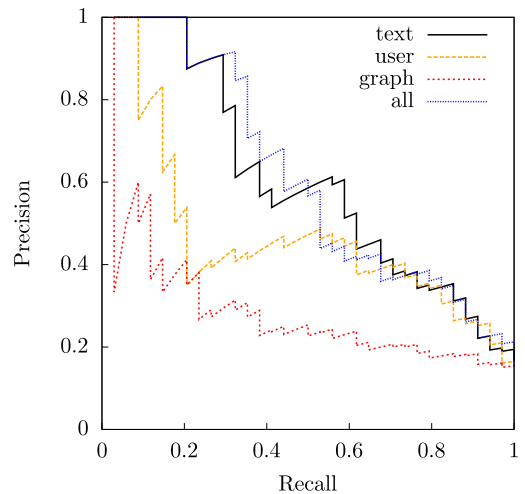


図 2 URL を含まないカスケードについての先頭 50 ユーザによる分類結果

Fig. 2 Classification results of information cascades without URLs using the first 50 users.

表 10 社会的影響力を持つ情報カスケードの細分類ごとの検知数

Table 10 The number of detected information cascades with social influence per each category.

分類	Text	User	Graph	All	総数
個人に対する意見	4	5	7	4	11
組織に対する意見	12	15	14	15	23
業界に対する意見	15	10	9	12	15
社会に対する意見	12	12	14	12	18
速報	7	10	10	6	20
注意喚起・デマ訂正	7	2	1	5	8
啓蒙	2	5	4	5	11

5.3.3 社会的影響力を持つ情報カスケードの細分類ごとの検知数

最後に、テキスト特徴量、グラフ特徴量、ユーザ特徴量の各特徴量が、3.4 節に述べた社会的影響力を持つ情報カスケードのどの細分類の項目の検知に寄与しているかを確認する。具体的には、各特徴量のみで分類する際の社会的影響力を持つ情報カスケード全体の再現率を、すべての特徴量を用いた場合に最良の F 値を取る際の再現率 (0.596) と同じになるよう分類の閾値を調整して、その際、どの細分類の項目の情報カスケードが正例として分類されているかを確認した。この結果を表 10 に示す。

この結果からテキスト特徴量は業界への意見や注意喚起・デマ訂正の検知に有効であるということが分かる。この原因としては業界に対する意見には「マスコミ」や「仕事」、注意喚起・デマ訂正としては「注意」「拡散」といった時期に依存しない特徴的な単語が多く見られることがあげられる。一方で、速報や啓蒙の項目については時事的な事柄が多く、特徴的な単語が学習データに含まれず分類することが困難だったと考えられる。またグラフ特徴量は個人への意見や、社会への意見の検知に有効である。この結果

から、個人批判や社会批判はマイクロブログ内で影響力の大きいユーザがつぶやいたりリツイートすることで爆発的に広まるため、その検知にはグラフ特徴量が有効的に働いたと考えられる。最後にユーザ特徴量は組織への意見、速報の検知に有効であることが分かる。この結果からは常に特定の組織を批判しているユーザや、事件の情報カスケードの情報を広めているユーザが存在することが読み取れる。

このように、社会的影響力を持つ情報カスケードの多様性に対処するには各特徴量を合わせて用いることが有効であることが確認できた。各特徴量ごとに検知に有効な社会的影響力を持つ情報カスケードは異なるため、ユーザが意図するカテゴリに有効な特徴量のみを組み合わせるというアプローチも有効であると考えられる。

6. まとめと今後の課題

本研究では社会的影響力を持つ情報カスケードの早期検知という新しいタスクを提案し、これを教師あり学習に基づく分類器を用いて解く手法を提案した。

社会的影響力を持つ情報カスケードの性質を明らかにするため、被験者実験により収集した社会的影響力のある情報カスケードについて、その内容を分類した。この知見に従い、本文に含まれる単語情報やカスケードのグラフ構造の特徴量、カスケードごとのユーザ特徴量を特徴量とした分類手法を提案した。実験では、社会的影響力を持つ情報カスケードの自動検知を行い、カスケードサイズ（リツイート数）が50の段階で、F値として0.576、さらに600の段階でF値0.709で検知が可能であることが示された。これはすべてのカスケードが影響力ありだとした場合（Baseline）のF値0.45と比較して顕著な改善である。

さらにテキスト特徴量の性能について詳細な分析を行った結果、テキスト特徴量はURLを含まないカスケードについてはよい検知性能を示していることが分かり、URLの中でも多くを占める画像によって社会的影響力を持つ情報カスケードによって性能が低下していることが分かった。

また、3.4節に述べた社会的影響力を持つ情報カスケードの細分類の項目ごとに分類性能を分析した結果、分類ごとに検知に有効に働く特徴量は異なることが判明した。そのため、社会的影響力を持つ情報カスケードの多様性に対処するには各特徴量を合わせて用いる本手法は有効であることが確認された。

今後の課題としては、訓練データの拡充、より小さなカスケードを対象とした実験、およびURLを含む情報カスケードに対する新しい特徴量の考案の3点があげられる。訓練データの拡充については、今回のデータセットは訓練セット1130、テストセット475とかなり小さいものであったため、検知精度の向上、および評価の信頼性向上のためにはより大きくすることが求められる。訓練データのサイズは検知精度の向上に直結するため、特に大規模化したい

ところではあるが、人手のアノテーションを行う限り大幅に大規模化することは困難である。そのためこれを解決するために自動で訓練データを獲得する方法を検討したい。2点目の小さなカスケードに対する実験については今回の実験は社会的影響力の有無による判別が主眼だったため、成長しきったカスケードのスナップショットをもとに検知を行った。しかし、広まらずに収束するような批判を早期に発見するというニーズも考えられるため、小さな情報カスケードに対する実験も試みたい。3点目のURLを含む情報カスケードに対する新しい特徴量としては、単純には画像特徴量の導入、外部記事へのリンクであればその記事のテキスト特徴量、また、新しくカスケードへのメンションのテキスト特徴量などが考えられる。

謝辞 本研究の一部はJSPS科研費25280111, 16H02905, 16K16109の助成を受けたものです。

参考文献

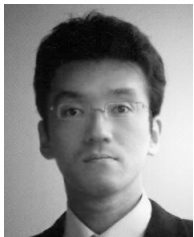
- [1] Allan, J., Carbonell, J., Doddington, G. et al.: Topic Detection and Tracking Pilot Study Final Report, *Proc. the DARPA Broadcast News Transcription and Understanding Workshop*, pp.194–218 (1998).
- [2] Bakshy, E., Hofman, J.M., Mason, W.A. and Watts, D.J.: Everyone's an Influencer: Quantifying Influence on Twitter, *Proc. WSDM*, New York, NY, USA, pp.65–74, ACM (2011).
- [3] Bikhchandani, S., Welch, I. and Hirshleifer, D.A.: A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades, *Proc. Journal of Political Economy*, pp.992–1026 (1992).
- [4] Blei, D.M. and Lafferty, J.D.: Dynamic topic models, *Proc. ICML*, pp.113–120 (2006).
- [5] Castillo, C., Mendoza, M. and Poblete, B.: Information credibility on twitter, *Proc. WWW*, pp.675–684 (2011).
- [6] Chen, P.-C., Lee, H.-M., Tyan, H.-R., Wu, J.-S. and Wei, T.-E.: Detecting spam on Twitter via message-passing based on retweet-relation, *Proc. TAAI*, pp.56–65 (2014).
- [7] Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M. and Leskovec, J.: Can cascades be predicted?, *Proc. WWW*, pp.925–936 (2014).
- [8] Fleiss, J.L.: Measuring nominal scale agreement among many raters, pp.378–382 (1971).
- [9] Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z. and Kellerer, W.: Outtweeting the twitterers-predicting information cascades in microblogs, *Proc. WOSM* (2010).
- [10] Gao, H., Chen, Y., Lee, K. et al.: Towards Online Spam Filtering in Social Networks, *Proc. NDSS* (2012).
- [11] Rattanakritnont, G., Toyoda, M. and M.K.: Characterizing Topic-Specific Hashtag Cascade in Twitter Based on Distributions of User Influence, *Proc. APWeb*, pp.735–742 (2012).
- [12] Landis, J.R. and Koch, G.G.: *The measurement of observer agreement for categorical data*, pp.159–174 (1977).
- [13] Ren, Z., Peetz, M.-h., Liang, S., Dolen, W.V. and Rijke, M.D.: Hierarchical Multi-Label Classification of Social Text Streams, *Proc. SIGIR*, pp.213–222 (2014).
- [14] Rill, S., Reinel, D., Scheidt, J. and Zicari, R.V.: PoliTw: Early detection of emerging political topics on twit-

- ter and the impact on concept-level sentiment analysis, *Proc. Knowledge-Based Systems*, pp.24-33 (2014).
- [15] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M.: Short text classification in twitter to improve information filtering, *Proc. SIGIR*, pp.841-842 (2010).
- [16] Tsur, O. and Rappoport, A.: What's in a Hashtag?: Content Based Prediction of the Spread of Ideas in Microblogging Communities, *Proc. WSDM*, pp.643-652 (2012).
- [17] Zhou, X. and Chen, L.: Event detection over twitter social media streams, *Proc. VLDB J*, pp.381-400 (2013).
- [18] 斎藤翔太, 富岡亮太, 山西健司: ソーシャルネットワークにおける長期間流行する話題の早期検出, 電子情報通信学会技術研究報告, IBISML, 情報論的学習理論と機械学習, Vol.111, No.480, pp.77-84 (2012).



川本 貴史

2014年東京大学工学部・電気電子工学科卒業。2016年同大学大学院情報理工学系研究科修士課程修了。現在楽天株式会社に勤務。



豊田 正史 (正会員)

東京大学生産技術研究所准教授。1994年東京工業大学理工学部情報科学科卒業。1996年同大学大学院情報理工学系研究科修士課程修了。1999年同大学院情報理工学研究科博士後期課程修了。博士(理学)。同年, 科学技術振興事業団計算科学技術研究員。ウェブマイニング, ユーザインタフェース, ビジュアルプログラミングに興味を持つ。ACM, IEEE CS, 日本ソフトウェア科学会各会員。



吉永 直樹

東京大学生産技術研究所准教授。2000年東京大学理学部情報科学科卒業。2002年同大学大学院理学系研究科修士課程修了。2005年同大学院情報理工学系研究科博士課程修了。博士(情報理工学)。2002年より2008年まで日本学術振興会特別研究員(DC1, PD)。2008年東京大学生産技術研究所特任研究員, 特任助教, 特任准教授, 情報通信研究機構主任研究員を経て現職。自然言語処理・計算言語学の研究に従事。

(担当編集委員 岡崎 直観)