

複数の国で生じた Twitter のトレンドの関係性の分析

Analysis of Relationships between Twitter Trends That Appeared in Different Countries

清水 将開

Masaharu Shimizu

法政大学情報科学部コンピュータ科学科

E-mail: masaharu.shimizu.5h@stu.hosei.ac.jp

Abstract

With the spread of smartphones, the number of users of social media such as social networking services and blogs is increasing, and connections between society and people are diversifying. This causes two problems. The first problem is that social trends change quickly and cannot be noticed in real time. The second problem is that such diversifying connections make international relationships unclear. To solve these problems, this paper is focused on "Twitter trend". A Twitter trend indicates a trend of tweets in a particular area with a specific keyword or a hashtag. This paper proposes a method for analyzing social trends by using Twitter trends to find relationships, characteristics, and trend propagation between countries. The proposed method consists of two steps. The first step uses Levenshtein distances to extract the same trend words. The second step compares the times when those words occurred. The data used is a 21-day trends of 17 countries that consist of Japan and 16 countries using English as the official language or a frequently used language. We conducted three experiments to confirm relationships, characteristics, and trend propagation between countries. The results show certain relationships, part of characteristics, and trend propagation among countries.

1. はじめに

近年、スマートフォンの普及に伴い、SNS やブログなどのソーシャルメディアの利用者が増加している。これらを背景に、社会や人とのつながりが多様化している。それに伴い、問題点として、流行の移り変わりが速く、今現在何が流行しているのかわからない点や、社会と人のつながりが多様化したことにより、国同士の特徴や関係性が不明瞭になっている点が挙げられる。

それらを解決するために、本研究では Twitter の「トレンド」に着目する。Twitter の機能の 1 つである「トレンド」は、特定のキーワードや、ハッシュタグと呼ばれる # を先頭につけたキーワードから、社会で今現在話題となっている事象や流行を、地域を指定して知ることができる機能である。本研究の目的は、国それぞれのトレンド

ドを時系列に比較し、国同士の関係性や特徴を分析することで社会動向を把握することである。本研究の新規性は、特に Twitter のトレンドを使用することにある。それによって、流行や国同士毎の特徴、関係性を分析する。一方、本研究の有用性は、国同士毎の関係性を分析することで、トレンドの流れを把握することが可能になることと、国同士の特徴を分析し把握することで、その国同士の社会動向を把握することが可能になることである。

本研究では提案手法を用いることで、国同士の関係性や特徴の一部、トレンドが Twitter 上で伝播したと考えられるものだけでなく、他のメディアを経由して伝播していることを確認することができた。次に、特定のワードを含むトレンドを抽出し、実際に仮定通りの結果なのか、確認と分析を行い、同時性のあるトレンドの特徴と関係性を確認することができた。時間経過で伝播するトレンドは、関係性や特徴の一部を確認することができた。これらの分析により、完全ではないが、社会動向を把握することが可能であると考えられる。

2. 関連研究

Twitter のデータを用いた研究は数多く行われているが、Twitter のトレンドを用いた研究はほとんど行われていない。Twitter のトレンドを用いた関連研究として、特定の地域に関するトレンドを評価する方法を提案している Albishry らの研究 [1]が挙げられる。Twitter を用いた関連研究として、都市に対する市民の関心の傾向を分析している安藤らの研究 [2]や、災害情報に対するツイートの伝播とその内容の感情を分析している三浦らの研究 [3]、コミュニティ内で導入された用語がコミュニティ外にどのように伝播するのかを分析している Golbeck らの研究 [4]、特定クエリを含む地理的位置付きツイートからネットワークを形成し、それがどのような要因によるものなのか分析する Zhang の研究 [5]が挙げられる。

3. 提案手法

国同士の関係性や特徴を分析するために、各国で生じた Twitter のトレンドを用いる。しかし、Twitter のトレンドは、指定した地域のトレンドなので方言や訛り、言語が異なるなどの理由で複数の国のトレンドを比較することや指定地点だけのトレンドでは、地域同士の関係性や特徴を分析することはできない。これらを解消するために、提案手法を用いる必要がある。

まず初めに、レーベンシュタイン距離を用いて、同一のトレンドの抽出を行う。次に、その同一トレンドの発生時刻やその量から国同士の関係性や特徴の有無を確認し、分析を行う。

3.1. 同一トレンドの抽出

同一な内容であるトレンドを抽出する手法として、レーベンシュタイン距離を使用する。これは、2つの文字列を比較し、片方の文字列を挿入、削除、置換の3つの操作でもう一方の文字列に置き換える最短手数を求める手法である。レーベンシュタイン距離は文字列が長いほど値が大きくなってしまふ可能性があるため、文字数が長い文字列の文字数で距離を割ることで、正規化し距離を一定にする。距離が0.25よりも小さいものを同一のトレンドとして抽出した。

3.2. 発生時刻の比較

同一であるトレンドの初期発生時を散布図として可視化し、同時性や発生時刻の偏りの有無を読み取る。トレンドの発生時刻は、初期発生時刻だけを比較しているので、後にもう1度トレンドになった場合の時刻は考慮していない。

4. 実験

4.1. 分析対象

本研究で使用したデータは、2019年12月3日13時00分から15分ごとに21日間、17箇国それぞれのTwitterトレンドである。17箇国の選定方法は、英語を公用語または日常でよく使用する国16箇国と比較対象として日本を加えた。データの総件数は、1,305,450件である。表1にそれぞれの国のトレンドの個数、「#」から始まる個数、その割合、新規トレンドの個数、新規トレンドの発生割合を表す。

表1. 各国のトレンド

国名	トレンド(個)	#から始まる(個)	割合(%)	新規トレンド(個)	割合(%)
オーストラリア	91,364	40,515	44.3	2,080	2.3
カナダ	98,947	39,823	40.2	2,685	2.7
アイルランド	88,911	40,508	45.6	1,300	1.5
ガーナ	41,567	20,270	48.8	738	1.8
イスラエル	31,562	16,217	51.4	1,075	3.4
ケニア	89,843	45,846	51.0	1,514	1.7
マレーシア	39,887	19,762	49.5	478	1.2
ナイジェリア	92,582	47,213	51.0	2,013	2.2
ニュージーランド	41,914	18,380	43.9	405	1.0
パキスタン	85,104	56,703	66.6	1,327	1.6
フィリピン	88,178	41,962	47.6	2,218	2.5
シンガポール	40,411	26,631	65.9	578	1.4
南アフリカ	92,408	43,673	47.3	1,858	2.0
イギリス	95,314	43,388	45.5	4,827	5.1
アメリカ	93,375	36,373	39.0	8,081	8.7
ベトナム	100,813	49,296	48.9	9,366	9.3
日本	93,270	34,784	37.3	6,963	7.5

各国に出現するトレンドの個数は、ベトナムが最多で100,813個で、イスラエルが最小で31,562個であり、3倍弱の差がある結果となり、国によって大きな差異がある結果となった。しかし、#から始まるトレンドの割合は、ほとんどの国で、約40から50%で国による差異はない結果となった。新規トレンドの個数は、イギリス、アメリカ、ベトナム、日本が高い結果となった。トレンドの個数については、利用者数が関係していると考えられる。利用者数が多いと様々な人が様々ツイートをするので、トレンドの移り変わりが早くなると考えられる。

4.2. 結果

4.2.1. 同一トレンドの抽出

3.1節に述べた手法で、各国それぞれの同一トレンドを抽出した。下記の表2は、アメリカと各国それぞれの結果を抜き出したものとなっている。全てのトレンドのレーベンシュタイン距離の平均、レーベンシュタイン距離が0.25よりも小さいトレンドの個数、その中で「#」から始まるものの個数、その割合を表している。

表2. USと同一のトレンド

国名	平均距離	0.25より小(個)	#から始まる(個)	#の割合(%)
オーストラリア	0.888	536	257	47.9
カナダ	0.885	1,366	613	44.9
アイルランド	0.889	241	91	37.8
ガーナ	0.887	389	177	45.5
イスラエル	0.886	503	251	49.9
ケニア	0.886	264	122	46.2
マレーシア	0.894	175	89	50.9
ナイジェリア	0.888	414	162	39.1
ニュージーランド	0.887	70	32	45.7
パキスタン	0.886	159	113	71.1
フィリピン	0.887	352	195	55.4
シンガポール	0.888	252	116	46.0
南アフリカ	0.888	344	141	41.0
イギリス	0.886	1,320	613	46.4
ベトナム	0.884	5,233	1,803	34.5
日本	0.982	87	37	42.5

レーベンシュタイン距離を用いたので、日本は言語が異なるため、平均距離が大きくなる結果となった。また、同一であると思われるトレンドの個数は、近隣国であるカナダ、イギリスとベトナムが高い結果となった。#から始まるトレンドの割合だが、各国それぞれの割合と大きな変化は見られず、40から50%前後である。

4.2.2. 発生時刻の比較

3.2節に述べた手法で、各国それぞれのトレンドの発生時刻を散布図として表現した。下記の図は、アメリカとカナダの比較をしたものである(図1)。対角線は、同時刻であることを示しており、対角線から離れたものは発生時刻に差異があるものである。表3は、アメリカと各国の平均時間を表したものとなっている。また、そのトレンドの個数を表4に表す。

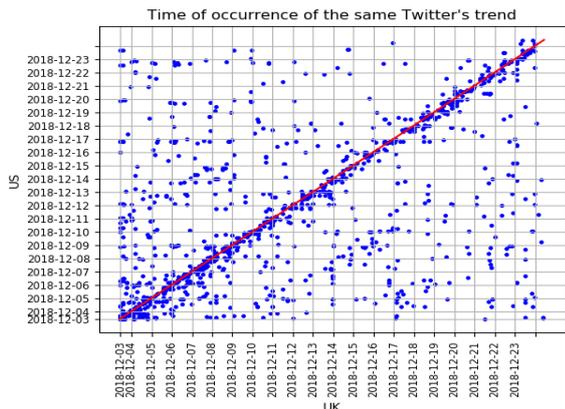


図 1. アメリカとイギリスの同一トレンドの発生時刻

表 3. アメリカと各国のトレンドの平均時間

国名	伝播平均時間(分)	USへの伝播時間(分)	USからの伝播時間(分)
オーストラリア	1,071	4,431	3,672
カナダ	1,312	4,653	3,480
アイルランド	-523	5,900	3,107
ガーナ	4,409	4,339	6,777
イスラエル	659	3,563	2,947
ケニア	259	6,209	4,191
マレーシア	-1,201	4,655	2,416
ナイジェリア	-659	2,563	2,947
ニュージーランド	-484	6,288	2,461
パキスタン	3,383	6,287	6,743
フィリピン	1,446	5,314	4,464
シンガポール	3,313	7,245	7,054
南アフリカ	306	6,433	4,121
イギリス	-10	5,248	551
ベトナム	1,566	551	5,099
日本	682	3,109	3,055

表 4. アメリカと各国のトレンドの発生時刻の個数

国名	USへ伝播したトレンド(個)	USから伝播したトレンド(個)	同時に出現したトレンド(個)
オーストラリア	168	259	9
カナダ	341	971	54
アイルランド	153	225	11
ガーナ	49	188	4
イスラエル	146	289	68
ケニア	99	163	2
マレーシア	81	69	25
ナイジェリア	180	221	13
ニュージーランド	23	45	2
パキスタン	41	118	0
フィリピン	105	239	8
シンガポール	61	181	10
南アフリカ	122	216	6
イギリス	576	708	36
ベトナム	631	617	538
日本	32	52	3

図 1 から対角線に多くのトレンドが集中していることが読み取れる。また、表 3、表 4 より多くのトレンドの発生時刻に差異があることが読み取れ、トレンドが伝播

していることが確認できる。また、表 3、表 4 よりアメリカは、多くのトレンドを伝播させている特徴があり、イギリスやケニア、南アフリカなどとは伝播時間が短くなっている。同一トレンドの個数よりカナダやイギリス、ベトナムと親密な関係性があると考えられる。

これらの結果から、全ての国々の特徴や関係性、トレンドの伝播を確認できるように、重み付き有向グラフを用いて可視化を行った(図 2)。ノードはそれぞれの国、ノードの大きさは各国のデータ数、エッジの向きは伝播の方向、エッジの重みは平均伝播時間、エッジの太さは同一トレンドの数となっている。

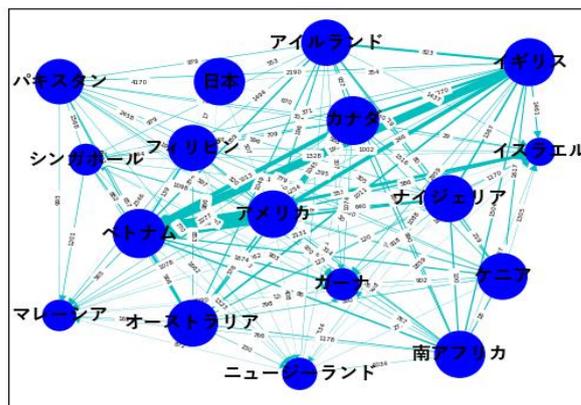


図 2. 各国のトレンドの伝播

図 2 からアメリカ、カナダ、イギリス、ベトナムの 4 箇国のエッジが太くなっていて、同一トレンドの数が多くトレンドが多く伝播していることが分かる。アメリカやカナダ、イギリスと近隣国で同一トレンドが伝播していることを考えると地理的に近いとトレンドが伝播していると考えられる。

5. 特定のワードによる分析

5.1. 時間経過で伝播するトレンド

4.2 節の結果から、国同士の関係性や特徴の一部、トレンドの伝播を確認することができた。しかし、4.2 節の結果からだけでは、具体的にどのようなトレンドが伝播しやすいのか、また、それぞれの国同士の関係性や特徴を完全に把握することはできない。そこで、これらを詳細に把握するために特定のワードによる分析を行った。

まず、各国それぞれでトレンドになり、時間経過で伝播するトレンドを分析するために、日付について着目した。日付は、各国の時間が異なるので、それについてのトレンドは、時間経過で伝播するトレンドであると仮定できる。そこで、「DAY」を含み 60 分以上して他国に伝播したトレンドを抽出し、分析を行った。

5.2. 同時性があるトレンド

次に、同時性のあるトレンドを分析するために、スポーツの試合や大規模な祭典についてのトレンドに着目した。スポーツの試合や大規模な祭典は、テレビ中継やネ

ット配信などのライブ配信により、国際的に同時にトレンドになる可能性が高く、同時性のあるトレンドであると仮定できる。そこで、「VS」を含むトレンドやチーム名を表す長さが3文字のトレンド、「2018」を含むトレンドかつ60分以下で他国に伝播したトレンドを抽出し、分析を行った。

5.3. 結果

5.3.1. 時間経過で伝播するトレンド

60分以上で伝播したトレンドの総数は、32,210個であった。その中で、条件に当てはまるものは、3,827個であった。また、条件外だったものは、268個あった。例としては、「#Mondaymotivation」や「#FridayFeeling」などであり、#から始まるものが多く、話題の共有をすることが目的であると考えられる。また、インドネシアで起きた津波に関するトレンドと考えられる「#Tsunami」のような国際的なニュースも存在した。これは、その国ではいち早くトレンドになり、その後実際には遠い国にも伝播したと考えられる。これらの結果から、時間経過で伝播するトレンドは、曜日や日付などの周期性があるトレンドが伝播しやすいと考えられる。また、その中でも短い時間で伝播しているものは、緊急性の高いものであると考えられる。つまり、時間経過で伝播するトレンドが多い国同士は、話題の共有する特徴があり、一定の周期で関わる関係性であると考えられる。

5.3.2. 同時性があるトレンド

60分以下で伝播したトレンドの総数は、5,516個であった。下記表5が、結果の一覧になっている。VSを含むトレンドは、スポーツに関するものが多く出現している結果となり、仮定通りの結果となった。しかし、一部のトレンドでは、映画の作品名であると考えられるトレンドも存在した。3文字のトレンドはなかったが、「#AUDvsIND」の様にVSを含み対戦カードを表すものや、「#GoPITs」の様にチームを応援していると考えられるものも存在した。チーム名を3文字で表すということは確認することができた。2018は、60分以上で伝播しているものがほとんどであった。祭典も周期的に行われているものであり、何度も行われているものは、時間経過で伝播するトレンドになると考えられる。

表5. 同時性があるトレンドの結果

条 件	当てはまる(個)	当てはまらない(個)
VS	121	236
3文字	0	0
2018	71	844
60分以下	5,516	

同時性があるトレンドの他の結果を確認すると人名や作品名が多い結果であった。これらの結果から、テレビ中継やネット配信の影響を受けたトレンドは、同時性があるトレンドになる可能性が高いと考えられる。つまり、同時性があるトレンドが多い国同士も同様な特徴があり、同じような事柄に関心がある関係性であると考えられる。

6. 議論

本研究の実験では、英語を公用語または、多く使用している国を対象に実験を行ったが、言語を変えた場合の調査もすることで、更に国同士の関係性や特徴が明確になると考えられる。表3の時間が4桁になってしまっている部分があるが、これはデータ量に対して経過時間が発散しているものの量が多かったと考えられる。伝播を考えずに最も近い時間で比べた場合、もっと妥当な数値となると考えられる。5節よりの結果を人名、スポーツ、イベントなど更に細かく分類することで、更に明確にどのトレンドが伝播することを確認することができ、国同士の関係性や特徴、伝播を更に明確にすることができる。また、それが可能ならば、次に伝播するトレンドを推測することができ、社会動向を明確に知ることができると考えられる。実験の問題点としては、言語が異なる場合、同一であるトレンドを抽出することができないので、言語が異なる国同士の関係性や特徴、トレンドの伝播は確認することができない。

7. おわりに

本研究の実験の結果から、国同士の関係性や特徴の一部、トレンドの伝播を確認することができた。しかし、特徴を完全には確認することができなかった。また、5節の結果から、同時性や時間経過で伝播するトレンドの特徴の一部を確認することができ、国同士の関係性や特徴を更に明確に確認することができた。また、本研究で使用した国以外のトレンドでも実験を行い、様々な国の結果を組み合わせることで、国同士の関係性や特徴を確認することができ、社会で次に話題となる社会動向を捉えることが可能になると考える。今後の発展としては、5節の分析をより詳細にすることで、言語が同じ場合のようなトレンドが次にトレンドになるか推測することができるようにと考えられる。

文 献

- [1] N. Albishry, T. Crick, T. Fagade and T. Tryfonas, "Popularity and Geospatial Spread of Trends on Twitter: A Middle Eastern Case Study," *Proc. ICCCI*, vol. 11055, pp. 167-177, 2018.
- [2] 安藤有生, 関洋平, "市民のツイートを用いた分散表現に基づく地名に対する市民の関心の傾向の可視化," *日本知能情報ファジィ学会誌*, vol. 30, no. 6, pp. 804-814, 2018.
- [3] 三浦麻子, 鳥海不二夫, 小森政嗣, 松村真宏, 平石界, "ソーシャルメディアにおける災害情報の伝播と感情: 東日本大震災に際する事例," *人工知能学会論文誌*, vol. 31, no. 1, pp. NFC-A_1-9, 2016.
- [4] J. Golbeck and C. Buntain, "This Paper is About Lexical Propagation on Twitter. H*ckin Smart. 12/10. Would Accept!," *Proc. ASONAM*, pp. 587-590, 2018.
- [5] X. Zhang, "Exploring the Country Co-occurrence Network in the Twittersphere at an International Economic Event," *Proc. Chinese National Conference on Social Media Processing*, vol. 774, pp. 308-318, 2017.