2P-1

Twitterにおけるツイート間の反応を考慮した話題分類法

塚田文哉† 鈴木徹也‡
芝浦工業大学 大学院理工学研究科 システム理工学専攻† 芝浦工業大学 システム理工学部 電子情報システム学科‡

1 はじめに

近年, ソーシャルネットワーキングサービス (SNS) という Web サービスが注目を浴び, 多くの人々が利用している. SNS とはインターネット上で社会的ネットワークを構築するサービスである. SNS の中でも Twitter [1] は 140 文字以内のメッセージを投稿できるミニブログである. Twitter への投稿はツイートと呼ばれる. 他のツイートへの返信はリプライという. そしてユーザ名を含むツイートはメンションと呼ばれる. ツイート内のハッシュ (#) で始まる文字列によって, 特定のトピックを明示することができる. その文字列をハッシュタグという.

Twitter における問題点として, ユーザが意図した話題のツイートを見つけることは困難であるということが挙げられる. ツイートの検索にはキーワード検索が主に利用されるのが原因である. また, 一つ一つのツイートの文章が短かいため, 読み手はそのツイートの内容だけでは意味を理解できないということがある. そのような場合, 文脈が理解の助けとなることがある.

そこで本研究では, Twitter におけるユーザのツイート内容理解支援を目的とした, ツイート間の反応関係に注目した話題分類法を提案する. ここでいう反応とはリプライ, リツイート, 引用を含むメンションである.

2 関連研究

2.1 時間的近さを考慮した話題構造マイニング

戸田らは, タイムスタンプを持つテキスト集合に対して話題構造マイニングを適用する手法を提案し, 時間類似度を考慮することが話題抽出及びクラスタリングの精度に及ぼす影響, およびその精度変化の要因について分析した [2]. 記事間の時間類似度には式 (1) を用いた.

$$TimeWeight(t) = T_0 * exp\left(-\frac{0.639}{t_{1/2}}t\right) \qquad (1)$$

t は二記事間のタイムスタンプの差, $t_{1/2}$ は時間類似度が 50% になるときのタイムスタンプの差 (半減期) を指す. T_0 は, タイムスタンプの差が 0 の場合の重みであり, 戸田らは 1 とした. この研究から TimeWeight に適切なパラメータを設定することで, 話題抽出の精度, クラスタリング精度が向上することが検証された.

2.2 文脈的つながりを考慮したツイートの抽出・提示手法の実現

青島らは, 特定の話題に関するツイートをまとめる為に, 単語の共起と時系列的な近さに基づく関連度に着目し, 単語間の関連度を算出する手法を提案した [3]. 時系列的な近さに基づく関連度の算出には式 (1) が用いられている.

2.3 データ圧縮による Twitter のツイート話題分類

西田らは, ツイートの話題分類を行う際にツイートの圧縮されやすさを応用した手法を提案した [4]. 提案手法では, データ圧縮を利用することで形態素解析に依存せず, 新語や口語・俗語が多く含まれるツイートを精度良く分類することができた.

A Topic Classification in Twitter considering Reaction among Tweets

Tsukada FUMIYA†, Tetsuya SUZUKI‡
†Division of Systems Engineering and Science, Guraduate School of Engineering and Science, Shibaura Institute of Technology
‡Department of Electronic Information Systems, College of Systems Engineering and Science, Shibaura Institute of Technology

3 反応の定義

次のリプライ, リツイート, 引用を含むメンションを本研究では反応と呼ぶことにする.

リプライ　Twitter における公式な返信機能によるツイート.

リツイート　他のユーザのツイートの再投稿.

引用を含むメンション　コメント, ユーザ名, そのユーザのツイートの先頭部分からなるツイート. 例えば, 図 1 の上部のツイートは, 下部のツイートの引用を含むメンションである.

図 1　引用を含むメンション

4 反応を考慮した話題分類

4.1 話題分類の流れ

ツイートの収集から話題分類までの方法を図 2 に示す. まず, システムではクローラを用いてツイートを収集する. 収集したツイートのトピックモデルの対象となるアカウントからトピックモデルを生成する. また, 収集したツイートとトピックとの関連度を求め適当なトピックにツイートを分類する.

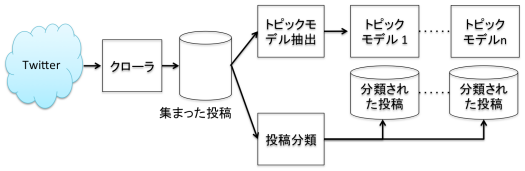

図 2　システム構成

4.2 クローラ

ツイートを収集するクローラを作成した. ツイートの収集は Twitter API を介して行う. 収集には以下の 3 つの手法を組み合わせて収集を行う.

Twitter Streaming API(sample)　ランダムサンプリングされたツイートを大量に収集する.

Twitter Streaming API(filter)　ニュース, 有名人などの他のユーザから反応されやすいアカウントのツイートを収集する. また, ニュース, 有名人のアカウントのフォロワーのツイートを収集する. ニュースと有名人のアカウントは, そのユーザのツイートのツイート数, 被メンション回数などを基に定期的に更新を行う.

Twitter REST API　Twitter Streaming API で取得したツイートについて, そのリプライ先, 引用したツイート, リツイート直後のそのユーザのツイートを収集する.

4.3 トピックモデル

本研究におけるトピックモデルは, ニュース記事とそのトピックに分類されたツイートとを連結したテキストである. トピックモデルの例を図 3 に示す. トピックモデルとなるツイートは Twtiter Streaming API(filter) を用いて収集する. ツイートはあらかじめ指定されたアカウントのツイートであり, 何らかの情報とその情報に対応するニュースサイトへの URL を含む.

4.4 ツイートのトピックへの反応度

トピック T とツイート d との反応度を示すために $ReactionWeight(T, d)$ を定義する.

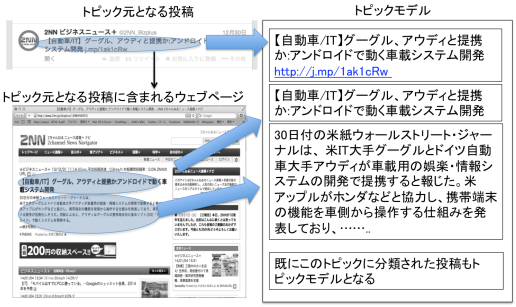

図3　トピックモデル

$ReactionWeight(T,d)$ は，ツイート d がトピック T に分類されているツイートへの反応であれば 1，そうでなければ 0 を取る．

4.5　ツイートのトピックへの関連度

話題分類における関連度を式 (2) 用いて算出する．$sim'(T,d)$ はトピック T とツイート d との文書間類似度を算出する．sim' は式 (3) で表され，λ はタイムスタンプの差の拡大による類似度の逓減の程度，NCD は正規化圧縮距離である [5]．$TimeWeight(T,d)$ は T と d とのタイムスタンプの差から時間類似度を算出する．α は $TimeWeight$ の重みである．

$$sim(T,d) = sim'(T,d) \ast((1-\alpha) \ast ReactionWeight(T,d) + \alpha \ast TimeWeight(T,d)) \quad (2)$$

$$sim'(T,d) = -\lambda \ast NCD(T,d) \quad (3)$$

4.6　ツイートの話題分類法

n 個のトピック $T_1,..,T_n$ があるとき，ツイート d を次のように分類する．まず T_1 から T_n の中でツイート d との関連度 sim を最大にするトピック T_i を 1 つ選択する．$sim(T_i,d)$ がある閾値以上ならばツイート d をトピック T_i に分類する．

5　実験

提案手法の効果を検証する為に実験を行った．

5.1　目的

ツイートのトピック分類において，反応関係を考慮した分類と考慮していない分類とではどのような違いが生ずるか確認する．

5.2　実験の条件

データセット　実験で用いるデータセットは 2013 年 12 月 8 日午前 5 時から翌 9 日午前 5 時までに投稿された 1,092,240 のツイート．
トピック数は 20 トピック．

分類手法　ベースライン式 (4) と提案手法式 (2) との二通りで話題分類を行った．α の値は 0.0 から 1.0 まで 0.1 刻みで変化させた．式 (3) の定数 λ は 0.913 とした．この値は，予備実験により α=0 としたときの式 (4) で分類したときの精度と再現率との F 値が最大となるように定めた．$TimeWeight(T,d)$ の定数 λ は半減期を 2 時間に設定し算出した．また，分類の際の閾値を各 α の値ごとに定めた．この閾値は予備実験により，精度がほぼ 100% となるような値を設定した．

$$sim(T,d) = sim'(T,d) \ast((1-\alpha) + \alpha \ast TimeWeight(T,d)) \quad (4)$$

関数 sim' の定数 λ:　0.913
$TimeWeight$ の定数 λ:　0.347
T:　トピックモデル
d:　ツイート
α:　TimeWeight の重み

5.3　実験結果

α の値を変化させた時の各手法での分類されたツイートの数を表 1 に示す．α の値が 0.0, 0.1, 0.2, 0.3, 0.4, 0.8 の時にベースラインが提案手法のツイート数を上回っている．特に α が 0.0 のとき，ベースラインと提案手法のツイート数はそれぞれ 772, 71 となり約 10 倍の差がある．

α が 0.0 の時のベースラインでの各トピックでのツイート数を表 2 に示す．ベースラインにおける各トピックのツイート数を比較してみると，トピック R におけるツイート数が多いことが分かる．トピック R を詳しく見てみると，トピック R とは関係のないツイートがほとんどであった．また，提案手法では反応であるツイートが見られた．

トピック R に関して各手法でのツイート数を表 3 に示す．α=0.0, 0.1, 0.2, 0.3, 0,4, 0,9 ではベースラインの方がツイート数が多い．しかしながら，ベースラインではトピックとは異なる話題のツイートが見られた．それに対し，α=0.5, 0.6, 0.7, 0.8, 1.0 では，提案手法の方がツイート数が同数，もしくはそれ以上であった．また，ツイート内容を詳しく見てみると，提案手法では反応であるツイートを含み，トピックに関する話題のツイートのみしか見られなかった．

表 1　α の変化による各種法のツイート総数

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ベースライン	772	106	90	75	72	63	62	58	69	55	49
提案手法	71	71	71	71	71	77	76	71	67	61	49

表 2　α=0.0 の時のベースラインにおけるトピック毎のツイート数

トピック	A	B	C	D	E	F	G	H	I	J
ツイート数	7	3	1	2	8	5	3	2	5	4
トピック	K	L	M	N	O	P	Q	R	S	T
ツイート数	1	3	20	1	6	7	33	654	6	1

表 3　トピック R における各種法でのツイート数

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ベースライン	654	7	7	5	5	5	3	3	6	7	6
提案手法	2	2	2	2	2	6	6	6	6	6	6

6　まとめ

ReactionWeight の効果を測る為にベースライン手法と提案手法の二手法を用いて評価実験を行った．評価実験から提案手法を用いると α の値を 0.5 から 0.7 の間で設定することで多くのツイートを含んだトピックを生成することができると考えられる．また，ReactionWeight を用いることで，反応であるツイートとトピックの話題に関係したツイートを正確に分類することができると考えられる．

参考文献

[1] Twitter. https://twitter.com/
[2] 戸田浩之, 北川博之, 藤村考, 片岡良治. 時間的近さを考慮した話題構造マイニング. DEWS2007 L6-4.
[3] 青島傳隼, 坂本翼, 横山昌平, 福田直樹, 石川博, 文脈的つながりを考慮したツイート群の効果的な抽出・提示手法の実現. 情報処理学会論文誌. データベース 6(2),61-84, 2013-03-29.
[4] 西田京介, 坂野遼平, 藤村考, 星出高秀. データ圧縮による Twitter のツイート話題分類. DEIM Forum 2011 A1-6.
[5] Ming Li, Xin Chen, Xin Li, Bin Ma, Paul M.B. Vitanyi. The Similarity Metric. IEEE TRANSACTION THEORY. VOL.50. NO 12. DECEMBER 2004.