ARG WI2 No.8, 2016
WI2-2016-03

Twitterの本文要素を用いたイベント視聴動向の推定

田中 千尋[†, a]　　若林 啓[††, b]

† 筑波大学図書館情報メディア研究科　†† 筑波大学図書館情報メディア系

a) *s1521627@u.tsukuba.ac.jp*　b) *kwakaba@slis.tsukuba.ac.jp*

概要　本研究では，ツイート情報を用いてテレビ番組の視聴者などの，実世界のイベントにおいて関わっている人の数を推定することを目的とする．本研究ではハッシュタグなどのつぶやきの本文要素とつぶやきを投稿したユーザ数，それらから求められる個々のユーザのつぶやき数などを用いることによって，より多くのユーザのつぶやきを対象としたイベント視聴動向の推定を行う手法を提案し，実験により有効性を検証した．その結果，ユーザ数とツイート数を用いることにより，ドラマの視聴率をある程度推測可能であることを明らかにした．

キーワード　Twitter, 視聴率推定, 回帰分析

1　はじめに

近年 Twitter をはじめとしたマイクロブログサービスの普及によって，多くのユーザが気軽に大量の情報を発信できるようになり，つぶやきから様々な情報を抽出する研究の発展が期待されている．その中でもイベント規模をソーシャルメディアの情報から推定できれば，当該イベントの関係者にとって，社会の反響を把握することが容易になり，視聴率などイベント視聴動向の測定にかかるコストを大幅に削減することができると考えられる．

これまでには，ツイート内で当該イベントに関連する単語を手がかりにした調査研究が行われているものの，手がかり語を専門家の知識を使わずに決定する必要があり，多様なイベントで網羅的に関連ツイートを収集することは困難である [1]．また，ツイートの位置情報を利用して，当該イベントの開催地で投稿を行ったユーザの数を利用して参加者数を推定する手法が提案されているが [2]，テレビ番組やオンライン上のイベントなどに適用することはできないという問題がある．

本研究では，Twitter において話題を明示的に表すハッシュタグを利用して，視聴動向を推定したい対象のイベントに関連していることが容易に推測可能なハッシュタグのみを手がかりとして用いる．与えられたハッシュタグのみを用いて関連ツイートを収集することで，多様なイベントの規模を推定する手法を提案する．

2　提案手法

本論文では，Twitter における投稿傾向からイベント規模の予測を行うことを考えるが，ツイート数はイベント規模に対して単純に比例するとは考えにくい．例として，表 1 にドラマ「息もできない夏」のイベント規模を表す指標としての視聴率と，ハッシュタグ「#息もできない夏」が付与されているツイート数，その異なりユーザ数を示した．他の放送回に対して第 8 回のツイート数が極端に大きくなっているにも関わらず，視聴率には大きな変化はなく，ツイート数のみを手がかりとして視聴率の予測を行うことは難しいと考えられる．しかし，異なりユーザ数もそれに応じて大きく増加していることから，本稿では異なりユーザ数の増加率の関係を手がかりにして予測精度を向上させることを検討する．

表 1 ドラマ「息もできない夏」の回ごとの推移

放送回	6	7	8	9	10	11
ツイート数	517	621	1038	585	450	503
ユーザ数	137	140	613	149	116	134
視聴率	8.1	10.6	8.2	11	7.8	8.6

2.1　データの収集

入力として,「容易に推測可能なハッシュタグ」1 つが与えられるとする．当該ハッシュタグは，その日の何らかのイベントに関連したものであると仮定される．このとき，アルゴリズムの出力として，入力のタグに関連付けられたイベントの規模を表す数値を求める問題を考える．これを推定するための知識として，あらかじめ以下のデータが利用可能であるとする．

- イベントの規模を数値化したデータ．このデータは,「当該イベントについての容易に推測可能なハッシュタグ」,「当該イベントの発生日」,「数値化された当該イベントの規模」の 3 つ組を 1 サンプルとしたデータセットとする．
- 各イベント発生日のツイートデータ

2.2　アルゴリズム

学習データの集合を $T = (t_1, ..., t_N)$，各学習データを $t_i = (t_i^{tag}, t_i^{date}, t_i^{scale})$ と表し，t_i^{tag} は当該イベントについての容易に推測可能なハッシュタグ，t_i^{date} は当

該イベントの発生日，t_i^{scale} は数値化された当該イベントの規模とする．各学習データ t_i について，日付が t_i^{date} のツイートデータからハッシュタグ t_i^{tag} の付与されたツイートを全て収集する．このツイート情報に基づいて，以下の 3 つの要素を説明変数として利用することを検討する．

- t_i^{tag} が出現したツイート数 $x_i^{numtweet}$
- t_i^{tag} を含むツイートを行ったユーザ数 $x_i^{numuser}$
- ユーザあたりの平均ツイート数 $x_i^{average} = \frac{x_i^{numtweet}}{x_i^{numuser}}$

これらの変数を用いた線形回帰モデルにより予測を行う．

3 実験

本実験では，イベント規模の推定に効果的な説明変数ついて明らかにすることを目的とする．実験に使用するデータは以下の通りである．

- イベント規模データ：2012 年の 7 月〜9 月のドラマ 15 本．各ドラマは 8 回から 12 回の放送があり，サンプル数は合計で 153 件である．
- ツイートデータ：各ドラマについて 1 つずつ設定した「容易に推測可能なハッシュタグ」が付与されたツイートデータ 62837 件．

実験では，視聴率を被説明変数とした回帰モデルを学習して，決定係数 R^2 を見ることで，相関のある結果が得られるかどうかを確認する．また，クロスバリデーションによる予測精度を，2 章で述べた 3 つの説明変数の有無全てについての組み合わせで構成される線形回帰モデルについて算出し，最も予測精度の高い説明変数の組み合わせを明らかにする．

4 結果

まず，それぞれの説明変数を用いて全データによる回帰モデルを学習し，その決定係数 R^2 を求めた．この結果を表 2 に示す．

表 2 回帰モデルごとの決定係数

回帰に用いた説明変数	R^2
$x_i^{numtweet}$ のみ	0.2424
$x_i^{numuser}$ のみ	0.1898
$x_i^{numuser}$，$x_i^{average}$	0.3246
$x_i^{numtweet}$，$x_i^{average}$	0.3482
$x_i^{numtweet}$，$x_i^{numuser}$，$x_i^{average}$	0.363

この結果から，候補である 3 つの説明変数を全て用いたときの決定係数が最も大きく，学習データに対して当てはまりのよいモデルを推定できているといえる．

クロスバリデーションにより，テストデータに対する直線予測に対する予測誤差を確認した結果を表 3 に示す．この結果からは，最良のモデルは「ツイート数」と「ユーザ数」のみを用いた場合であることが分かる．

表 3 用いた説明変数と予測誤差

$x_i^{numtweet}$	$x_i^{average}$	$x_i^{numuser}$	予測誤差
使用	未使用	使用	0.09297
使用	使用	未使用	0.09580
使用	使用	使用	0.09702
未使用	使用	使用	0.09714
使用	未使用	未使用	0.09866

ツイート数とユーザ数を用いて重みの学習を行った結果は，視聴率の予測数 y とすると，以下の式となった．

$$y = 7.677398 + 0.007765x_i^{numtweet} - 0.018771x_i^{numuser}$$

これより視聴率に対してユーザ数とは負の相関が，ツイート数とは正の相関があることが示された．ユーザ数が増えればツイート数は当然増えるが，ユーザ数の増加に比べてツイート数の増加が小さい場合は，個々のユーザが継続的にツイートをしていないためにすぐに見るのをやめてしまっており，ユーザ数の増加に比べてツイート数の増加が大きい場合は，個々のユーザが継続的にツイートをしており視聴者数が多いと考えられる．

平均ツイート数の説明変数は，ツイート数とユーザ数から間接的に求められるため，この 2 つの説明変数を用いている場合には，平均ツイート数の説明変数を加えることの効果は小さいことが分かった．

5 結論

本研究ではユーザ数とツイート数を用いることにより，ドラマの視聴率をある程度推測可能であることを明らかにした．今後の課題としてはドラマ以外のイベントへの適用の検証と，容易に推測可能なハッシュタグが付いていないイベント関連ツイートを網羅的に収集する手法の適用により予測精度を向上を目指すことが挙げられる．

謝辞

本研究の一部は，NII 戦略研究公募型共同研究，JSPS 科研費（課題番号 16H02904）および筑波大学図書館情報メディア系プロジェクト研究の助成によって行われた．

参考文献

[1] ビデオリサーチ社:視聴率と Twitter の関係解析 —「Twitter TV エコー」データ分析より—, 2015.

[2]Federico Botta, Helen Susannah Moat, Tobias Preis：Quantifying crowd size with mobile phone and Twitter data，Royal Society Open Science，No. 5, 2015.