ソーシャルメディアの投稿状況に基づく
イベント参加動向の推定に関する研究

筑波大学

図書館情報メディア研究科

2017年3月

田中 千尋

目 次

図 目 次

表目次

第1章　序論

1.1　背景

近年 Twitter をはじめとしたマイクロブログサービスの普及 [1] 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 [2][3] 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することができれば，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．

しかし，イベントの動向の測定を行うためにはそれぞれのイベントごとにふさわしい手がかり語を用いて検索を行い，集計をする必要がある．手掛かり語は対象がテキストで構成された論文や Web ページなどのオブジェクトであれば当該のオブジェクトからキーワードを抽出することが可能であると考えられるが [4]，一方イベントを対象に考えた場合，どのような手掛かり語を用いるのかはイベントの分野によって変動し，検索者の裁量によって検索結果に大きな差が出てしまうこととなる．

1.2　イベント参加者数の推定

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 [5]．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 [6]，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．

1.3 本論文の構成

本論文の構成は以下の通りである．第 2 章では本研究と関係のある研究について述べ，本研究がこの分野でどのような役割を果たすのかを明らかにする．第 3 章ではハッシュタグを用いてイベント規模の推定を行う手法についての提案を行う．第 4 章では提案手法の有効性を検証するための実験を行い，その結果に対して考察を行う．第 5 章では本論文のまとめを行う．

第2章　関連研究

ここではイベント規模の推定を行う際に，どのような変数が効果的であるのか検証を行う．

SNS から情報を抽出し，実世界の情報の推定を行う研究として，Minnich らによる研究が挙げられる [7]．この研究では複数のレビューサイトから情報の合併，比較，評価を行う手法が提案されている．この手法では複数のサイトのレビューを用いることによって，個別で評価を行った時と比べて約 7 倍の問題のあるホテルを検出することが出来た．また，三つのサイト全てに掲載されているホテルの内 20%程度が低スコアのホテルであった．

Hovy の研究によると [8]，現在メタデータが付与されているデータは少なく，メタデータが付与された大規模データが不足しているとされている．Hovy は作者のメタデータを付与して書き込みを行うレビューサイトを立ち上げ，研究を行った．メタデータを用いることによって情報抽出の制度の向上を図ることができるが，メタデータを大規模データに付与するためには非常に大きな労力，資金が必要となる．そのため，メタデータを用いない手法の精度の向上が必要とされている．

ビデオリサーチ社の研究によると [5]，視聴率とツイート数には相関関係があるということを示した．この相関関係はユーザの年齢によって変化し，ユーザが若年であれば相関は高く，高齢者であれば相関は低いということを示した．しかし，網羅的にツイートを収集するためには手がかり語を専門家の知識を用いて決定する必要があり，他の分野で同様の実験を行うことは難しい．

また，Botta の研究では [6]，Twitter の位置情報とスタジアムの来場者数の相関を調べることによって，Twitter の位置情報から来場者数を高精度で推測することができると示した．しかし，位置情報を用いているため，オンライン上のイベントや，テレビ番組などには適用することができない．

第3章　イベント規模の推定手法

3.1　イベント規模の推定に効果的な変数の検証

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しない場合もあると考えられる．例として，表 3.1 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 3.1: 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

3.1.1　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

3.1.2　容易に推測可能なハッシュタグの条件

容易に推測可能なハッシュタグは，対象ドラマのタイトルそのまま，略称，もじりなど様々なパターンが考えうるが，本論文では容易に予測可能なハッシュタグの条件を「ドラマのタイトルから記号と副題，シリーズ番号を取り除いたもの」と定義する．

この定義に則ったハッシュタグの例として，ドラマ「ハンチョウ　警視庁安積班　5」の容易に推測可能なハッシュタグ,「#ハンチョウ」．ドラマ「チーム・バチスタ 3　アリアドネの弾丸」の容易に推測可能なハッシュタグ,「#チームバチスタ」などが挙げられる．

3.1.3　アルゴリズム

学習データの集合を $T = (t_1,\ldots, t_N)$ ，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 [9].

3.2　ハッシュタグクラスタリングを用いた手法

また,「容易に推測可能なハッシュタグ」を用いることは検索者の裁量によって検索の精度が左右されてしまう危険性があり，さらに表記揺れによって，そのドラマを示しているツイートをしているが，検ハッシュタグには別の表現を行った場合を拾い損ねてしまう可能性がある．例としては,「パパドル！」というタイトルのドラマを対象としてハッシュタグを検索する場合，上記の手法では「#パパドル」というハッシュタグで検索を行うこととなるが，実際には本来のタイトルをそのままハッシュタグにした「#パパドル!」, タイトルをローマ字表記にした「#papadoru」なども考えられる．このような検索漏れを減らすためにハッシュタグクラスタリングを用いた手法の実験を行う．

ハッシュタグクラスタリングは井上らによって提案された手法で [10]，ハッシュタグを TF-IDF によってベクトル化し，それらを k-means[11] によってクラスタリングすることによって類似した目的で使用されたハッシュタグの抽出を行うという手法である．このハッシュタグクラスタリングを用いることによって，表記揺れによって取りこぼしてしまったハッシュタグをより網羅的に抽出できると考え，実験を行う．

3.3　一週間のデータを用いた予測

上記の二つの実験では当日のツイートデータを用いることによって視聴率の予測を行い，そのデータが一致しているかどうか判断を行う．それに対して，対象とする放送回の過去一週間のデータを用いて予測をすることによって，未来のデータの予測を行った際の精度についての実験を行う．

第4章　実験と評価

本章では，提案手法の有効性を評価するための実験を行う．実験で用いるイベント集合は特定の期間に放送されたドラマのデータで，その視聴率をイベント規模として推定を行う．

ドラマの集合として，以下の 2 種類を実験に用いることで，放送期間の違いによる結果の変化の有無について調べる．ただし，4 月から放送していたドラマ,「たぶらかし　代行女優業・マキ」は日にちをまたがって放送がされていたため，今回は実験対象から除外する．

- 2012 年の 4 月〜6 月のドラマ
- 2012 年の 7 月〜9 月のドラマ

それぞれのドラマについて 3 章で述べたように，容易に推測可能なハッシュタグを用いる方法，ハッシュタグクラスタリングで得られるハッシュタグ集合を用いる方法，容易に推測可能なハッシュタグを用いて一週間前までのツイートを収集する方法で，それぞれツイートデータを収集し，線形回帰による視聴率の予測を行う．

また，リツイート（RT）の影響を調べるため，RT であるツイートを含む場合と含まない場合の比較についても実験を行う．

4.1　イベント規模の推定

4.1.1　7 月 ∼ 9 月期のデータ

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の内リツイートではないもの 62,837 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，3.1 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．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 4.1 に示す．

表 4.1: 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．

ツイート数，ユーザ数と視聴率の関係を図 4.1，図 4.2 に示す．この場合ツイート数，ユーザ数ともに低い値に集まり，また数が増えた際には大きく値が跳ね上がっているため，両対数を適用することによって回帰の精度の向上を試みる．両対数をとった場合のツイート数，ユーザ数と視聴率の関係を図 4.3，図 4.4 に示す．

クロスバリデーションにより，テストデータの両対数をとった値の直線予測に対する予測誤差を確認した結果を表 4.2 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 4.2: 用いた説明変数と予測誤差

$\ln x_i^{numtweet}$	$\ln x_i^{average}$	$\ln x_i^{numuser}$	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．

$$y = 7.\,677398 + 0.\,007765x_i^{numtweet} - 0.\,018771x_i^{numuser} \tag{4.1}$$

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．

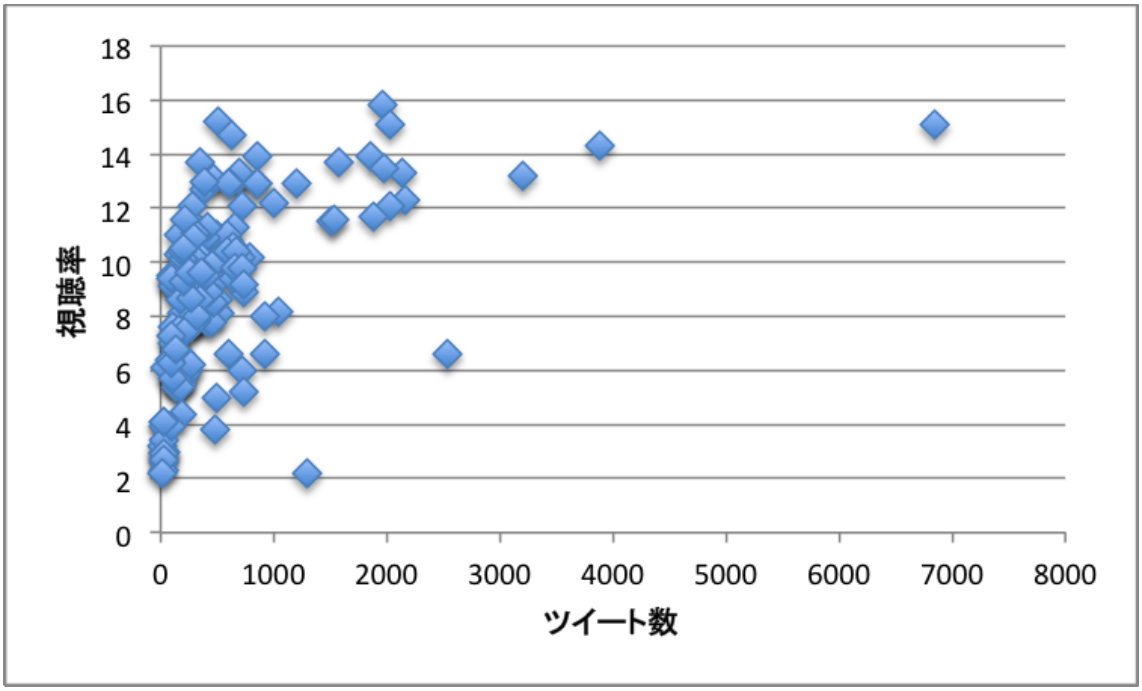

図 4.1: tweet 数と視聴率の関係

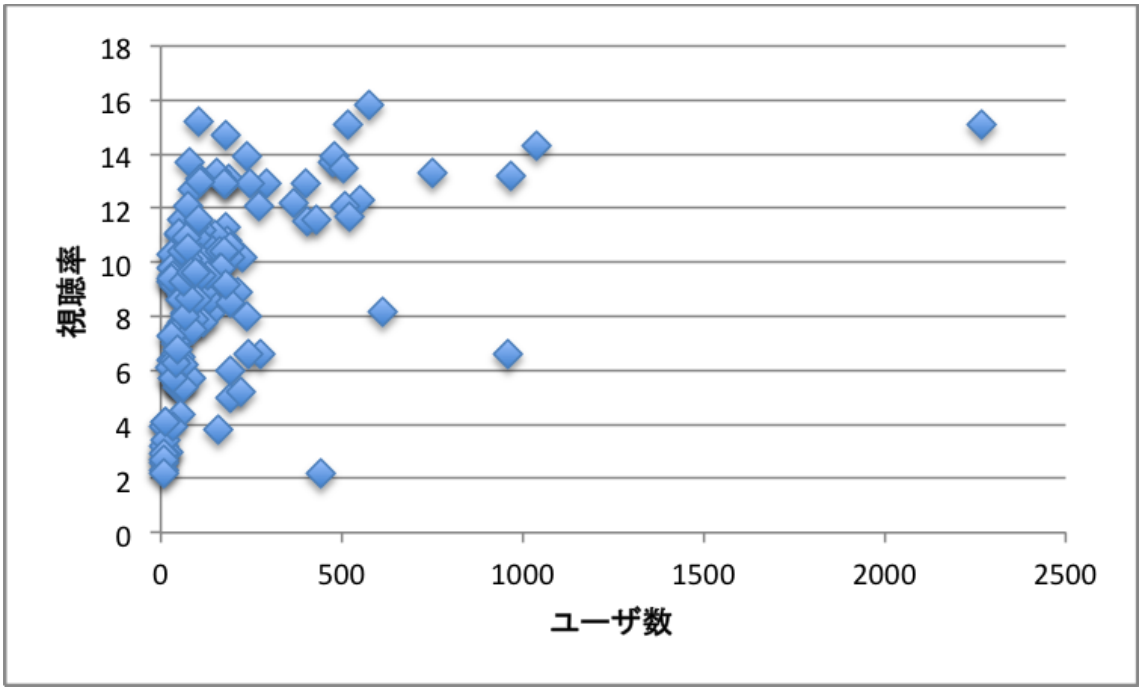

図 4.2: user 数と視聴率の関係

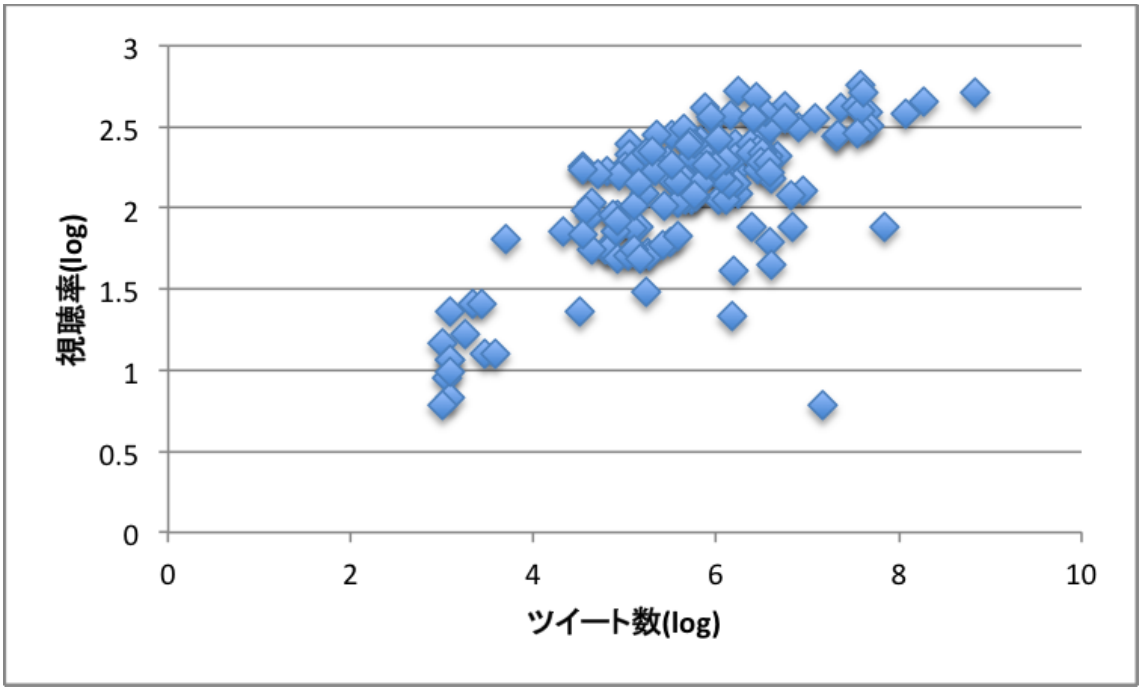

図 4.3: 両対数をとった tweet 数と視聴率の関係

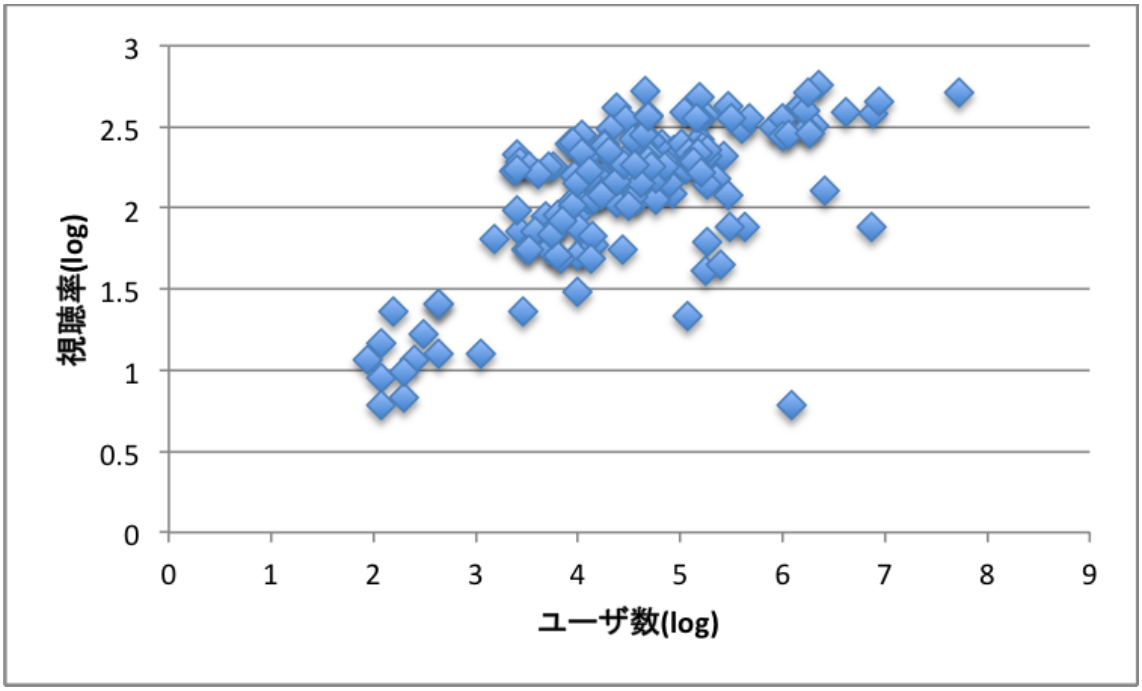

図 4.4: 両対数をとった user 数と視聴率の関係

4.1.2　4 月 ~ 6 月期のデータ (RT を含まない場合)

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 4 月〜6 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 151 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の内リツイートを含まないもの 134,809 件．

この実験では平均ツイート数を用いずに実験を行った．クロスバリデーションにより，テストデータの両対数をとった値の直線予測に対する予測誤差を確認した結果を表 4.3 に示す．

表 4.3: 4 月 ~ 6 月期のデータ (RT を含まない場合)

$\ln x_i^{numtweet}$	$\ln x_i^{numuser}$	予測誤差
未使用	使用	0.1210
使用	未使用	0.1214
使用	使用	0.1263

この場合は「ユーザ数」のみを使用した場合が最も精度が高くなっている．

4.1.3　4 月 ~ 6 月期のデータ (RT を含む場合)

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 4 月〜6 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 151 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 170,954 件．

この実験では上記の実験のデータに加え，それらのツイートのリツイートも集計した．クロスバリデーションにより，テストデータの両対数をとった値の直線予測に対する予測誤差を確認した結果を表 4.4 に示す．

表 4.4: 4 月 ~ 6 月期のデータ (RT を含む場合)

$\ln x_i^{numtweet}$	$\ln x_i^{numuser}$	予測誤差
使用	未使用	0.1212
未使用	使用	0.1221
使用	使用	0.1256

この場合は「ツイート数」のみを使用した場合が最も精度が高く，逆に「ツイート数」と「ユーザ数」の両方を用いた場合に最も低い精度となっている．リツイートを集計したことによって 36,145 ツイートもデータが増えたにもかかわらず，リツイートを集計しなかった場合と予測誤差の値はほぼ変わらなかった．これはリツイートが一定の確率で発生し，全体の比率に影響を与えていないのではないかと推測することができる．

4.2 ハッシュタグクラスタリングを用いた手法

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 4 月〜6 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 151 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」と同じクラスタに含まれるハッシュタグが付与された放送開始時間の前後 12 時間以内につぶやかれたツイートデータ 199,722 件．

この実験では 4 月〜6 月のドラマに関係したツイートのハッシュタグに対してクラスタリングを行うことによって対象のドラマに関するハッシュタグをより網羅的に収集し，さらに最も多くのつぶやきが行われると考えられる放送開始時間の前後 12 時間に対象を絞ることによって抽出するツイートを増やし，精度の向上を試みた．

その結果，抽出したツイート数は 28,768 ツイート増加したものの，その大半がツイートを収集する時間帯を変えたことによるもので，ハッシュタグクラスタリングによって追加されたハッシュタグはドラマ,「ハンチョウ 5」のハッシュタグ「#ハンチョウ」と同じクラスタに含まれる「#hancho」のみであった．

クロスバリデーションにより，テストデータの両対数をとった値の直線予測に対する予測誤差を確認した結果を表 4.5 に示す． この結果からは，最良のモデルは「ツイート数」のみを用いた場合であることが分かる．

表 4.5: ハッシュタグクラスタリングを用いたデータ

$\ln x_i^{numtweet}$	$\ln x_i^{numuser}$	予測誤差
使用	未使用	0.1237
未使用	使用	0.1238
使用	使用	0.1283

4.3 一週間のデータを用いた予測

この実験ではドラマの放送日一週間前から放送前日までの一週間のデータを用いて未来のドラマの視聴率の予測を行った場合の精度を確かめる．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 4 月〜6 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 151 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」が付与された放送日の前一週間以内につぶやかれたツイートデータ 191,702 件．

クロスバリデーションにより，テストデータの両対数をとった値の直線予測に対する予測誤差を確認した結果を表 4.6 に示す． この結果からは，最良のモデルは「ツイート数」のみを用いた場合であることが分かる．

表 4.6: 未来の視聴率の予測精度

$\ln x_i^{numtweet}$	$\ln x_i^{numuser}$	予測誤差
使用	未使用	0.1297
未使用	使用	0.1311
使用	使用	0.1343

今回の実験では未来のデータの予測であったため，精度の低下が予想されたが，当日一日のデータを用いて予測を行った 場合と比較しても予測誤差の値が 0.0085 ほど低下したのみに留まった．

4.4 考察

4 月〜6 月期のデータではツイート数，ユーザ数の二つの変数を説明変数とした時に最も予測誤差が小さくなったのに対して，7 月〜9 月期のデータでは，ツイート数のみを使用した時に最も予測誤差が小さくなった．これは，実験結果がドラマの特徴に依存するという側面があるということが原因の一つであると考えることができる．例えば 3 章で例に挙げた「息もできない夏」であるが，8 話でツイート数が大きく伸び，話題になった．このような突然話題になる振る舞いをする特殊なデータの影響を受けたと考えられる．

また，データ量が不十分である可能性が示唆されるという側面も考えられる．より大量のイベントをデータとして用いて再度検証を行うことが今後の課題であると言える．

今回の実験ではハッシュタグクラスタリングの有効性を確認することができなかった．これは，大量のハッシュタグをクラスタリングするためには計算コストが高いため，本実験では文字数が 10 文字以内のハッシュタグで，かつ収集期間に 30 回以上出現しているハッシュタグのみを対象としてクラスタリングを行った．そのため容易に推測可能なハッシュタグと同じクラスタに入りうるハッシュタグの多くがはじかれてしまったのではないかと考えられる．

また，今回の実験ではハッシュタグを使ったツイートのみを対象として実験を行ったが，井上らの手法ではハッシュタグのついていないツイートに対してハッシュタグクラスタリングを推薦する手法についても言及しており，そのような手法を適用することも検討する余地があると考えられる．

第5章　結論

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用いることによって位置情報やユーザの情報などのメタデータが付与されていないデータに対してでもデータの抽出を行い，未来のデータの予測を行うことのできる手法の提案を行った．

3章ではドラマの視聴率がツイート数とは比例せず，視聴率の増加に対して非常に大きな変動を見せることがあるという例を示し，ドラマ視聴率に異なりユーザ数が可能であるのではないかという可能性について述べた．また，ハッシュタグクラスタリングを用いることによってより網羅的なハッシュタグの収集を行う手法の提案を行った．

4章では3章で提案した様々なパターンについて実験を行った．2012年7月～9月期のデータを用いて線形回帰を行うことによってドラマの視聴率を被説明変数とする式を示した．他にも，同じ時期のデータをリツイートを含む場合と含まない場合のふた通りで収集して回帰を行い，比較した．その結果双方で大きな差は見られず，リツイートは一定の確率で起きるのではないかということが示唆された．また，ハッシュタグクラスタリングを用いた手法を用い実験を行ってみたところ，ハッシュタグクラスタリングによって拾うことの出来たハッシュタグは非常に限定的であった．さらに，今までの実験は主にドラマ放送日当日のデータを用いてその日の視聴率の推定を行うといったものであったが，放送前一週間のデータを用いての予測を行った，その結果当日のデータを用いた場合とほぼ同等の予測誤差の値を示した．この結果から提案手法が放送前のソーシャルメディアの投稿状況から視聴率を予測するということにも利用可能であるということが示唆される．

謝辞

本研究を進めるにあたって何度も手伝い，助けてくださった若林研究室の星川祐人さん，久保田豊久さん，ありがとうございました．また，主指導を引き受けてくださり，研究についてアドバイスもして頂いた鈴木伸崇先生．研究を行っていく中で非常に助けになりました，ありがとうございました．また，論文の審査を二つ返事で快く引き受けてくださった手塚太郎先生，ありがとうございました．そして非常に手のかかる生徒であった私を最後まで指導してくださった若林啓先生には言葉にできないほど感謝をしております．本当にありがとうございました．

参考文献

[1] Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng. Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis.2007, p56-65.

[2] 伊集 竜之, 遠藤 聡志, 山田 孝治, 當間 愛晃, 赤嶺 有平. Twitter ユーザの年齢層推定のための有効な素性の検討 (一般:データ分析と応用). インテリジェントシステム・シンポジウム講演論文集.2014(24), p121-124.

[3] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors Proceedings of the 19th international conference on World wide web,2010, p851-860.

[4] 白井 宏和, 春原 将寿, 中村 勝一, 横山 節雄, 宮寺 庸造. 論文構成要素に着目した論文間関係把握支援システムの開発電子情報通信学会技術研究報告. 2009, Vol.108, No.470, p23-28.

[5] 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.

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

[7] Amanda J. Minnich, Nikan Chavoshi, Abdullah Mueen, Shuang Luan, Michalis Faloutsos. TrueView: Harnessing the Power of Multiple Review Sites. Proceedings of the 24th International Conference on World Wide Web. 2015, p 787-797.

[8] Dirk Hovy, Anders Johannsen, Anders Sogaard. User review sites as a resource for large-scale sociolinguistic studies. Proceedings of the 24th International Conference on World Wide Web. 2015, p 452-461.

[9] 根岸 拓郎, 藤田 悟 . 携帯端末のセンサ値を用いた多変量解析による歩幅推定. 第 77 回全国大会講演論文集. 2015, p299 - 300.

[10] 井上 優作, 若林 啓. 表記の多様性を考慮したハッシュタグ推薦. DEIM Forum 2016. 2016, B6-5.

[11] Douglas Steinley, Michael J. Brusco. Initializing k- means Batch Clustering: A Critical Evaluation of Several Techniques. Journal of Classification. 2007, Vol.24, No.1, p99-121.