社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

Twitter における実況書き込み検出手法の検討

小林 尊志† 野田 雅文† 出口 大輔† 高橋 友和‡ 井手 一郎† 村瀬 洋†

†名古屋大学大学院情報科学研究科 〒464-8601 名古屋市千種区不老町 1
‡岐阜聖徳学園大学経済情報学部 〒500-8288 岐阜県岐阜市中鶉 1-38

E-mail: †{tkobayashi, mnoda}@murase.m.is.nagoya-u.ac.jp, †{ddeguchi, ide, murase}@is.nagoya-u.ac.jp,
‡ttakahashi@gifu.shotoku.ac.jp

あらまし マイクロブログサービス Twitter では，実際にスポーツを観戦したり TV 番組を視聴したりしながらリアルタイムに書き込む“実況書き込み”が増加している．本報告では，実際に観戦・視聴していないユーザの書き込みなど，他の様々な書き込みの中からこれを検出する手法を提案する．実験では TV 番組において「番組によらない情報」と「番組固有の情報」に注目して単語の出現頻度を学習して実況書き込みの検出を行った．

キーワード Twitter，実況，テキスト解析

Detection of On-the-spot Comments for Twitter

Takashi KOBAYASHI†, Masafumi NODA†, Daisuke DEGUCHI†, Tomokazu TAKAHASHI‡,
Ichiro IDE† and Hiroshi MURASE†

†Graduate School of Information Science, Nagoya University 1 Furocho, Chikusa-ku Nagoya, 464-8601 Japan
‡Faculty of Economics and Information, Gifu Shotoku Gakuen University 1-38 Nakauzura, Gifu, 500-8288 Japan

E-mail: †{tkobayashi, mnoda}@murase.m.is.nagoya-u.ac.jp, †{ddeguchi, ide, murase}@is.nagoya-u.ac.jp,
‡ttakahashi@gifu.shotoku.ac.jp

Abstract On the social networking and micro-blogging service “Twitter”, there has been increasing “on-the-spot comments” while watching a sport game at a stadium or watching TV at home. We aim at the detection of these on-the-spot comments among other comments on Twitter. This paper reports an experiment to detect on-the-spot comments for TV programs, analyzing the frequency of occurrence of words by focusing on both program independent and dependent information.

Keyword Twitter, On-the-spot comments, Text analysis

1. はじめに

ウェブ上のコンテンツから様々な評判情報を抽出する研究が盛んに行われている．藤村ら[1]は，Web 上の文書から商品などの評判情報を抽出する手法を提案している．また，宮森ら[2]は TV 番組に対する実況チャットの書き込みから評判情報を抽出し，これと番組映像を関連付ける報告をしている．

一方，近年ウェブ上で爆発的に普及しつつあるサービスとしてマイクロブログが挙げられる．その代表例である Twitter†1 は世界で 1 億 1 千万人のユーザが利用している．Twitter の書き込みは「ユーザ名」と「書き込み（最大 140 字）」から構成され，字数制限による気楽さにより携帯端末からの書き込みも容易なため，リアルタイム性が高い情報交換ツールとして活用されている．Twitter の書き込みから映画の評判を解析する報告[3]や TV 番組への評価を解析するサービス†2 があり，評判解析に用いる対象としても注目されている．

本研究では評判解析に用いるために，ユーザがイベントを観覧している(「観覧」)かしていない(「不観覧」)かを Twitter の書き込み内容から自動で識別することを目的とする．ここで「イベント」とは，スポーツの観戦や TV 番組の視聴など多数の人が同一の事象にする状況を表す．またこの状況を「観覧」と呼ぶことにする．本報告ではイベントを観覧しているユーザによる実況書き込みの検出手法を提案する．

†1 マイクロブログ Twitter：http://twitter.com/
†2 盛り上がりを視覚化「テレビジン」：http://tvz.in/

2. 実況書き込み検出手法

イベント開催時間中に投稿された書き込みのうち，関連すると思われるキーワードを本文に含むものを“実況書き込み候補”とする．これらを Twitter の API[†3] を利用して自動的に収集する．収集された実況書き込み候補は「投稿者名」「本文」「投稿時間」を含んでいる．これらを次の 2 段階で絞り込む．

(1) イベントによらない情報を用いたラベル付け

イベント開始直後の書き込みは「なう」や「はじまた」などイベント内容に依存しないものが多い．そこで様々なイベント開始直後の書き込みからイベント内容に依存しない表現を抽出し，イベント開始直後の単語の出現頻度を学習することで識別器を構築する．

注目するイベントの開始直後の書き込みを，構築された識別器で識別し，開始直後に書き込みをしたユーザに対して「観覧」「不観覧」のラベル付けを行う．

(2) イベント固有の情報による書き込みの絞り込み

（1）により「観覧」「不観覧」のラベルが付いたユーザのイベント開始直後以外の書き込みを解析することで，イベント固有の情報を抽出して検出器を構築する．具体的には，（1）においてラベル付けされたユーザの書き込みについて開催時間全体にわたる単語の出現頻度を学習して検出器を構築する．

構築した検出器により実況書き込みの検出を行う．

3 実験

多数のユーザが同時に観覧するイベントの例として，TV ドラマの実況書き込みの検出実験を行った．

3.1 番組によらない情報によるラベル付け精度

2010 年 1 月～3 月に放送された TV ドラマ 5 タイトル（各 40 分～54 分）の放送 1 回分に対して，放送開始 5 分間に書き込まれた 614 件の実況書き込み候補を用い，ユーザへのラベル付け実験を行った．

各実況候補を形態素解析し，単語の出現頻度上位 200 語を要素とする 200 次元の特徴ベクトルで表した．ユーザへのラベル付け精度を 3 fold cross-validation で評価した．識別器には SVM を用い，予め人手でラベル付けしたものを正解とした．その結果，平均 90.1% の精度で付与できることを確認した．

3.2. 番組固有の情報による絞り込み精度

TV ドラマ 1 タイトル（54 分）の書き込み 815 件について，番組固有の情報による絞り込み実験を行った．また，イベントによらない情報を用いた 1 段階目のラベル付けは，人手で行ったものを用いた．

以下の 3 つの手法により，出現頻度上位 500 語を要素とする 500 次元の特徴ベクトルを生成した．

†3 Twitter API：http://apiwiki.twitter.com/

表 1　実験結果：TV 番組実況の検出精度

	手法 1	手法 2	手法 3
検出率（%）	73.7	76.3	79.9

表 2　検出失敗例：False-positive

例 1	イライラする。やっぱライアーゲーム見るのやめよう。
例 2	ライアーゲーム真剣に見ちゃったけど、先週見てないからわからな…そうでもなかった。

・手法 1：名詞のみ

・手法 2：名詞，動詞，助動詞

・手法 3：名詞，動詞，助動詞，形容詞，形容動詞

各手法の検出率を表 1 に示す．

3 fold cross-validation で評価した結果，手法 3 が最も良く検出できることを確認した．ここで検出器には SVM を用い，人手で付与したラベルを正解とした．

3.3. 考察

表 2 中に検出失敗例を示す．これらはともに実際は「観覧」しているのに「不観覧」として誤検出された例である．失敗例 1 は実際には視聴しながらの書き込みであるが，「やめよう」とあるため，誤検出されたと考えられる．失敗例 2 は「見ちゃった」と「見てない」の両方が含まれている複雑な書き込みであるため，誤検出されたと考えられる．

複文や長文で構成された複雑な書き込みは，単なる単語の出現頻度による検出は難しく，文単位の解析や単語の出現順序や組合せを考慮した検出器を構築する必要があると考えられる．

4 むすび

本報告では，Twitter からの実況書き込み検出手法を提案した．TV ドラマを対象として書き込み内容の単語出現頻度を学習し，実際にイベントを観覧しながらの書き込みを検出した．実験の結果，提案手法の有効性が確認できた．

今後は観覧者の書き込みからの評判抽出や，不観覧者の書き込みから不観覧の原因抽出を検討している．

謝辞　本研究の一部は科研費による．

文　献

[1] 藤村滋，豊田正史，喜連川優，“文の構造を考慮した評判抽出手法，” 電子情報通信学会第 16 回データ工学ワークショップ，6C-i8，Feb. 2005.

[2] 宮森恒，中村聡志，田中克己，“番組実況チャットに基づく視聴者視点を利用した放送番組のビュー生成，” 電子情報通信学会第 16 回データ工学ワークショップ，4B-i9，Feb. 2005.

[3] S. Asur and B.A. Huberman, “Predicting the future with social media (informal publication),” ArXiv e-prints, 1003.5699, Mar. 2010.