1K3-4

スポーツ大会における コンテンツに着目したリツイート行動の要因分析

A Contents-based Analysis of Retweet Behavior in Sports Events

小林竜也　　尾崎知伸
Tatsuya Kobayashi　Tomonobu Ozaki

日本大学文理学部
College of Humanities and Sciences, Nihon University

This paper reports the results of contents-based analysis of twitter messages on 2013 World Championships in Athletics. In the analysis, we assess what kind of content in twitter messages, *e.g.* athlete names, technical terms, emoticon and frequent terms, has positive or negative effect for the retweet by using regression analysis, decision trees and propensity score matching.

1. はじめに

近年，SNS やマイクロブログが爆発的に普及している．マイクロブログの一つである Twitter は，最大 140 文字の記事を投稿・閲覧するコミュニケーションサービスであり，手軽な情報交換ツールとして幅広く利用されている．Twitter 上で受け取ったメッセージを転送する行為をリツイートと呼ぶ．リツイートは，利用者が受け取ったメッセージに強い関心を持った場合に行われる行為であり，リツイート行動を分析することで，利用者の興味等を推測することが期待できる．また，リツイートに強い影響を与える要因を分析することは，情報伝播を促進する，またはいわゆる炎上を回避するという点でも有効であり，種々の観点から研究が行われている．例えば文献 [1] では，コンテンツ（ツイート本文）に着目し，ハッシュタグの有無や，感嘆符・疑問符の有無，顔文字や感情語の有無が，どの程度リツイートに影響を与えるかを調査・分析している．

本研究では，文献 [1] と同様コンテンツに着目し，世界規模のスポーツ大会の一つである第 14 回世界陸上競技選手権大会に関するツイート群を対象に，リツイート要因の分析を行った．

2. 分析データの準備

本研究で利用したデータは，第 14 回世界陸上競技選手権大会（世界陸上モスクワ 2013）に関するツイートである．大会の開催期間中にハッシュタグ “世界陸上”，“ 世陸”，“seriku” を用い，Twitter 社が公開している Streaming API 経由でツイートの取集を行った．その後，十分に時間をおいてから収集したツイートのリツイートの有無やその回数を確認した．収集されたツイートの総数は 67,839 であり，そのうちの約 40%にあたる 27,268 ツイートがリツイートされたという結果となった．

次に，本研究で分析の対象とした属性を示す．実際の分析では，各属性の有無が，リツイートにどのような影響を与えるのかを調べることになる．

第 n 頻出語： ツイート本文を対象に形態素解析を行うことで抽出した，名詞頻出上位 100 語それぞれ．

ハイパーリンク： URL と判断できる文字列．

ユーザネーム： ユーザネームと判断できる文字列．

連絡先: 尾崎知伸，日本大学 文理学部 情報科学科，〒 156-8550 東京都世田谷区桜上水 3-25-40, tozaki@chs.nihon-u.ac.jp

表 1: 各属性を含むツイート数

総ツイート数	67,839	リツイート数	27,268
ハイパーリンク	9,193	顔文字	439
ユーザーネーム	27,504	陸上用語	49,968
感情語（喜）	747	競技名	23,660
感情語（怒）	13	決勝出場選手名	31,387
感情語（哀）	78	日本人選手名	19,165
感情語（恐）	846		

顔文字： 判定には，1059 語からなる独自辞書を利用した．

感情語： 感情表現辞典 [2] で示される 4 つの感情（喜，怒，哀，恐）それぞれに関する感情語．判定には，[2] に基づき，喜 268 語，怒 217 語，哀 246 語，恐 163 語からなる辞書を作成・利用した．

陸上用語： 日本陸上競技連盟公式が提供している陸上競技用語集 [*1] に含まれる用語（157 語）．

競技名： 公式名称から略称，経験者からの呼称など，独自に準備した辞書（86 語）を用いて判定を行った．

決勝進出選手名： TBS 世界陸上公式サイト [*2] から引用し，苗字のみを登録した辞書（380 語）を利用した．

日本人選手名： 苗字と名前をそれぞれ登録した辞書（88 語）を利用した．なお，愛称は省略している．

表 1 に，収集したツイートが，各属性をどの程度含むのかをまとめる．表より，陸上に関する専門用語は多くのツイートに含まれている一方，感情語や顔文字を含むツイート数が極端に少ないことが分かる．

3. 結果と考察

本研究では，リツイートの有無に関する要因を分析するために，ロジスティック回帰分析及び決定木を用いた．一方，リツイート回数に対しては，線形回帰分析と回帰木及び傾向スコア

∗1 http://www.jaaf.or.jp/international/glossary.pdf
∗2 http://www.tbs.co.jp/seriku/result/

マッチングを適用した．各分析において，目的変数をリツイートの有無もしくは回数とし，前章で示した属性を説明変数としてモデル化を行っている．以下，各分析実験の結果を示す．

ロジスティック回帰分析の結果

リツイートされたか否かを目的変数としたロジスティック回帰分析において，絶対値の大きな係数を持つ説明変数（属性）を表 2 にまとめる．表より，文字列 “daijapan”（為末大，男子元陸上競技選手のアカウント名）を含む場合にリツイートされやすいという結果となった．また，文字列 “西塔” と “拓己” は，競歩の西塔拓己選手を表すと考えられるが，苗字と名前でその影響が逆転するという結果となっている．

表 2: ロジスティック回帰分析の結果

4.03	daijapan	1.47	良子	1.29	ウクライナ
1.70	拓己	1.45	今日	1.28	km
-1.68	4 × 100m	1.39	心配	1.26	本日
1.48	mr	-1.35	西塔	1.26	通過

線形回帰分析の結果

リツイート回数を目的変数とした線形回帰分析の結果を表 3 に示す．表中では，係数の絶対値が大きなもののみを示している．表より，関係の強い属性は，ロジスティック回帰の場合とよく似ていることが分かる．また，“高瀬” は高瀬慧選手を，“良子” は木崎良子選手をそれぞれ表すと考えられ，選手名もリツイートに貢献していることが読み取れる．一方で，“4 × 100m” に関しては，ロジスティック回帰分析と正負が逆転しており，どの様に解釈すべきか判断が難しい結果となった．

表 3: 線形回帰分析の結果

526.9	心配	105.8	daijapan	74.8	最終
148.5	4 × 100m	101.3	良子	60.0	高瀬
124.6	niigata	78.2	今日	-56.4	銅
120.0	ウクライナ	76.2	競歩	56.2	金メダル

決定木分析の結果

目的変数 (クラス) をリツイートの有無とした場合の決定木による分析結果を図 1 に示す．結果より，ツイート本文に陸上用語や決勝進出選手名が含まれる場合，リツイートされる可能性が高くなることが示唆されている．また，“男子”，“daijapan”，“女子” といった文字列が強く関係しており，これらの文字列を一つも含まない場合，リツイートされる可能性が低くなることが分かる．

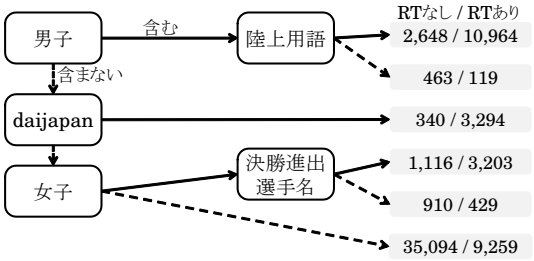

図 1: 決定木による分析結果

回帰木分析の結果

目的変数をリツイート数とした場合の回帰木による分析結果を図 2 に示す．結果より，“アメリカ” や “獲得” など新たな頻出語がいくつか現れているが，強い影響を持つ要因としては，概ねこれまでの分析と大差ない結果となった．

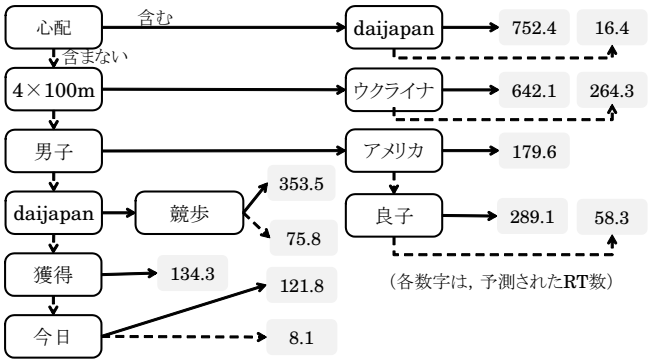

図 2: 回帰木による分析結果

傾向スコア分析の結果

目的変数をリツイート数とした場合の傾向スコアマッチング [3] による分析結果を表 4 に示す．表中の値は，その属性が本文に含まれる場合に見込まれるリツイートの増加数である．短距離選手の名前（高瀬，飯塚，アリソン）や競技名（4 × 100m）に加え，“恐れ” に関する感情語が，大きな影響を与えていることが分かる．

表 4: 傾向スコア分析の結果

681.0	心配	417.4	ウクライナ	272.6	速報
622.9	恐	384.2	高瀬	223.0	飯塚
426.4	4 × 100m	306.9	アリソン	221.1	daijapan

以上，5 種の分析全体を通じ，専門用語よりも頻出語の影響が大きいことが示唆された．また，日本人が活躍する競技は正負限らずその影響が大きいことも示唆された．その一方で，リツイートの有無と回数で，影響の正負が逆転する場合も存在し，更なる検証が必要である．

4. まとめ

本論文では，世界陸上に関するツイートを対象に，ツイート本文に出現する種々の要素がリツイートに対しどのような影響を与えるかを分析した．

今後の課題としては，投稿時間やフォロワーネットワークなど，コンテンツ以外の要素に着目したリツイート分析があげられる．また，陸上競技以外のスポーツや，スポーツ以外の分野を対象とし，それぞれの結果を比較することで，より分野に特化した要因を明らかにすることも重要な課題の一つである．

参考文献

[1] N. Naveed, T. Gottron, J. Kunegis and A. Che Alhadi: Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter, *Proc.* of the 3rd International Conference on Web Science (WebSci’11), 2011.

[2] 中村明：「感情表現辞典」，東京堂出版，1993.

[3] 星野崇宏：調査観察データの統計科学—因果推論・選択バイアス・データ融合，岩波書店，2009.