3M-9

テレビ番組のダイジェスト自動生成のための Twitter 分析

品田 勇一† 宮川 大志† 羽山 徹彩‡
金沢工業大学情報学部† 金沢工業大学工学部‡

1. はじめに

テレビ番組のダイジェクトはその内容を端的に伝えたり，視聴するかどうかの指標を与えたりするために有用であるが，その作成には多くの手間や労力，或いは良い感性を要するため，その自動化が求められている．従来研究の多くは音声・動画や字幕を利用して，その盛り上がりや内容を分析することで，ダイジェクト自動生成技術が開発されてきた [1)2)]．しかしながら，利用データが音声の強弱，特定の物体画像認識，および一時的な文字列データといった単純な特徴データを扱っているために，各番組内容に含まれる多様な観点をダイジェクトに反映することが難しかった．

近年，テレビ番組を見ながら，実時間にその様子やコメントを Twitter 投稿する人が増えてきた [3)]．我々はそのような膨大なデータをもとに，テレビ番組の盛り上がりや要点を判断するとともに，様々な観点から纏めた多様性を持ったダイジェクト自動生成が可能であると考えている．しかしながら，これまでテレビ番組に関する Twitter に対し，そのような目的をもとに分析がされておらず，ダイジェスト自動生成に利用可能かどうかも，ほとんど議論されてこなかった．

そこで本研究ではテレビ番組のダイジェスト自動生成を実現するために，テレビ番組に対して Tweet された Twitter データを分析したので，その結果について報告する．

2. Twitter データからテレビ番組ダイジェストの生成のために

テレビ番組のダイジェクトを生成するためには，番組内容の主な出来事とその標識，および観点による分類が可能であることが求められる．それらを Twitter データから行うためには，以下の点を調査する必要がある．

1. Tweet 頻度の時系列において，テレビ番組の主な出来事が反映されているか
2. それぞれの主な出来事の内容が Tweet 内容から，判断することが可能か
3. TV 番組での出来事を Tweet 内容から，観点による分類が可能か

Twitter Data Analysis for Automatic-Generating System of TV-program Digest Based on Twitter Data
† Kanazawa Institute of Science

1. に関しては， Tweet 頻度を時系列に並べ，その頻度が高い時間と低い時間の番組内容の出来事を確認する．2. に関しては TV 番組の出来事における Tweet を分析し，その内容に関するキーワードを確認する．3. に関しては Tweet を観点に基づき分類し，それぞれの時系列結果をもとに番組の出来事を分類する．

さらに，以下の点についての分析も試みる．

4. TV 番組を視聴しながら Twitter を利用しているユーザの特徴

3. 分析

3.1 対象データ

今回の分析対象データは，2013 年 11 月 16 日に放送されたサッカー国際親善試合(日本 VS オランダ)とした．Tweet データは，サッカー関係の Twitter ユーザとそのフォロワーの 44 万ユーザから事前に収集されたデータのなかで，放送時間 2 時間中の 376,656 件の Tweet である．またユーザ数は 51,565 であり，Tweet 数上位ユーザとその Tweet 数は図 1 のとおりである．

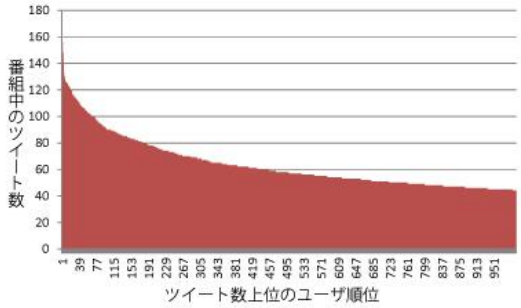

図 1 Tweet 数上位ユーザの番組での Tweet 数

3.2 結果

Tweet 頻度時系列と番組の出来事との関係

Tweet 頻度時系列に対し，番組の出来事の注釈を加えたグラフを図 2 に示す．

この番組内容であるサッカーでは主な出来事として，4 つの得点シーンが含まれており，それらが Tweet 頻度時系列に反映されていることがわかる．また実際の時間と Tweet の反映時間にはほぼ時差がなく，頻度が高まる傾向にある時点で，各イベントが発生していることを確認した．また全体として，試合の前半よりも後半の方が Tweet 数が増加していることがわかる．

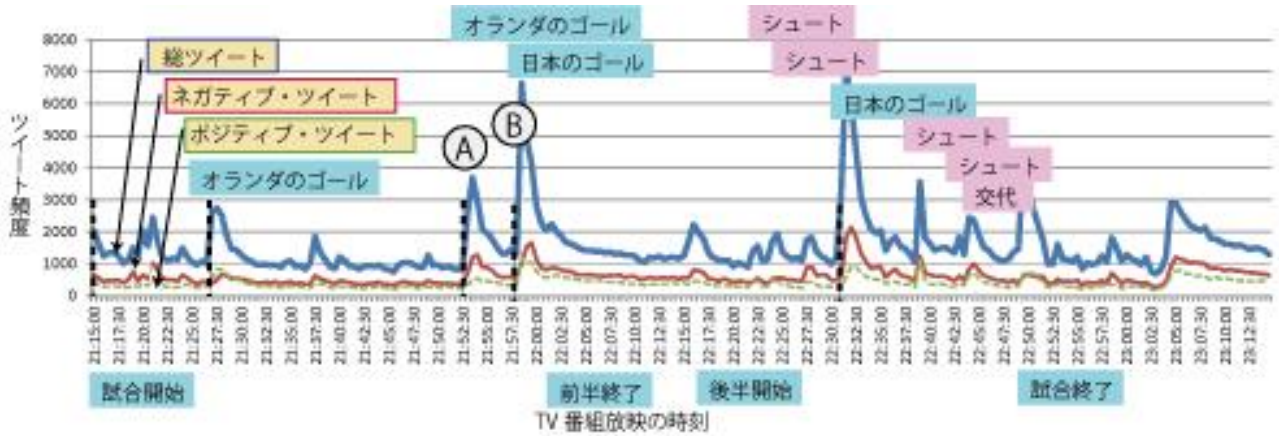

図 2 TV 番組放送(サッカー, 日本 VS オランダ)におけるツイート頻度の時系列

主な出来事とその時点での Tweet 内容

図 2 の得点シーン A および B で Tweet に含まれる単語出現数を確認した．その頻度上位 10 の単語を表 1 に示す．いずれも得点シーンであるが，シュートを決めた選手の名前の頻度が圧倒的に高いことがわかる．そのため，Tweet にはその時点での内容が含まれているといえる．

表 1 各シーン時の Tweet に含まれる頻出単語

順位	シーンA		シーンB	
	単語	出現頻度	単語	出現頻度
1	ロッベン	1890	大迫	7273
2	うまい	673	おお	2712
3	ロッペン	670	半端	2212
4	オランダ	663	ああ	1343
5	パス	356	いい	1014
6	ああ	351	松木	984
7	シュート	337	長谷部	785
8	すごい	321	点	784
9	さすが	314	ゴール	749
10	サイド	306	日本	577

出来事への Tweet 内容の観点による分類

Tweet をポジティブとネガティブの内容に分類し，それぞれの観点ごとに Tweet 頻度の時系列を調べた．そのグラフを図 2 に示す．Tweet とのポジティブ/ネガティブの判断は，先行研究 [4)5)] の評価表現を用いて，分類した．シーン A と B を比較し，オランダの得点シーン A の方が日本の得点シーン B よりもネガティブの Tweet の増加傾向が大きいことがわかる．そのため，Tweet 内容にはユーザの観点が含まれているといえる．

ユーザの分類

TV 視聴しながら Twitter するユーザを分類した．手順としては Tweet 数上位 100 ユーザに対して，Tweet 数，リツイートの割合，ハッシュタグの割合，ポジティブ・ツイートの割合，ネガティブ・ツイートの割合をスケーリングし，クラスター分析を行った．クラスター分析にはウォード法が用いられた．その結果を，表 2 に示す．

表 2 クラスター分析で分類したグループの特徴

グループ(ユーザ数)	平均 tweet 数	平均ハッシュタグ使用割合	平均リツイート割合	平均ポジティブ割合	平均ネガティブ割合
G1(24)	63.08	0.04	0.54	0.41	0.25
G2(13)	60.85	0.12	0.90	0.42	0.20
G3(30)	66.40	0.03	0.29	0.33	0.27
G4(24)	60.33	0.02	0.09	0.34	0.22
G5(9)	57.89	0.61	0.37	0.42	0.27

各グループにおいて，平均 tweet 数にほとんど違いがなかったものの，リツイート割合が高いグループ(G1, G2)，ハッシュタグ使用割合が高いグループ(G5)，タグをほとんど使用しないグループ(G4)，ポジティブ/ネガティブの Tweet 割合が同じグループ(G3)などの特徴が確認された．

4. おわりに

本研究では TV 番組ダイジェスト自動生成に，Twitter が利用できることを，実際のデータを分析することで確認した．

参考文献

1) 山本ら，サッカー映像のシーン自動解析の研究，通学技報．PRMU,104(573)，pp73-78 (2005)．
2) J.Wang, et al.: “Sport Highlight Dectection from Keyword Sequences using HMM,” Proc.ICME ’04, vol.1, pp.599-602, 2004.
3) テレビとソーシャルメディアの関係性，ネットエイジア株式会社，http://www.mobile-research.jp/.
4) 小林ら，意見抽出のための評価表現の収集．自然言語処理，Vol.12, No.3, pp.203-222, 2005.
5) 東山ら，述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第 14 回年次大会論文集，pp.584-587, 2008.