5S-03

バースト検出に基づく要約手法の検討

赤木文彦† 小城卓也‡ 田邊祐貴‡ 藤井章博‡
法政大学大学院理工学研究科† 法政大学 理工学部‡

1. はじめに

近年 Twitter では，テレビ番組に対する感想や内容を投稿するために用いられ，特定のイベントに関して投稿された tweet は”イベントストリーム”と呼ばれている．しかし，イベントストリームは移り変わりが頻繁であり，ユーザが関心のある tweet を閲覧し，イベントストリーム内の全ての内容を理解することは容易ではない．イベントストリームでは，視聴者の関心の高まりにより tweet 数が急激に上昇する現象がある．この現象は”バースト”と呼ばれている．

本研究ではバーストを解析し，要約を生成することを目的とし，過去の tweet からユーザが視聴番組の内容を理解することに対しての支援を目指す．特に，2014 年 7 月 13 日にかけて行われたブラジル開催 FIFA ワールドカップの日本代表戦に焦点をあてる．情報の配信が最も多く行われた日本対ギリシャ戦の推移を以下に示す．図 1 は 1 分毎の tweet1 数の推移である．

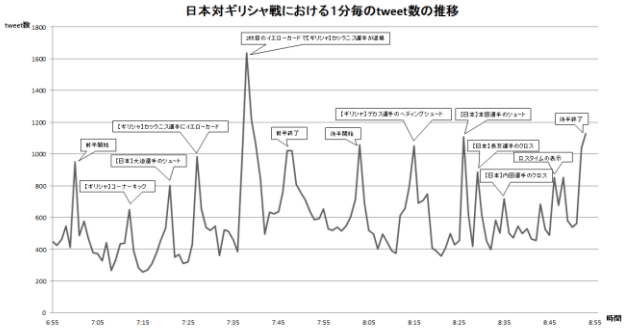

図 1 日本対ギリシャ戦(1 分毎の tweet 数の推移)

本研究では，日本対ギリシャ戦に着目して述べる．

2. 関連研究

久保ら[2]はイベント内で要約を生成する際の提案手法をバースト検出とバースト内の tweet 集合からの要約生成の 2 つに分けた要約システムを開発した．

坂本ら[4]は時間によってトピックの移り変わりを捉える目的で要約を生成し，Twitter 上からキーワードを抽出する際に，バーストの断続性に着目したリアルタイムバースト検出を行った．

3. 提案手法

3.1 リアルタイムバースト検出

本研究では，蝦名ら[1]の提案したリアルタイムバースト検出手法を用いる．蝦名らの手法は，一定時間毎に tweet 数の上昇を観測し，バーストを解析するのではなく，tweet 毎にバーストの解析を行うことができ，短時間に大量の tweet が行われら際に，高速に処理を行うことができるため，リアルタイムバースト検出に適したアルゴリズムである．蝦名らの提案した手法では，Aggregation Pyramid と呼ばれるセルをデータとするプラミッド構造を用いる．図 2 に Aggregation Pyramid を示す．セルのデータ構造のレベル 0 は N 個のセルを持つ．上層のセルは下層のセルのデータを統合した情報を持つ．

合計到着間隔(gaps)，到着時間(arrt)，間隔個数(gapn)をセルの各データとする．各セルを同条件で比較するために 1 つのセル内の到着間隔の 1 つあたりの平均値を求める平均到着関数を以下のように定義し，avg(c(h, t))と avg(c(N-1, t-1-h))を比較する．

$$avg(c(h,t)) = \frac{c(h,t).gaps}{c(h,t).gapn}$$

バーストを判定するパラメータ β (0< β <1)を用いて以下の条件を満たすとき．バーストが発生していると定義する．

$$avg(c(h,t)) \leq \beta \times avg(c(N-1, t-1-h))$$

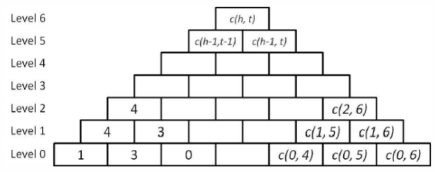

図 2Aggregation Pyramid

ピラミッド構造を決定する N，バースト判定係数 β，セルの最小ウィンドウサイズ W_{min} は，経験的に N = 60，β = 0.80，W_{min} = 3000(ミリ秒)とした．

† Graduate School of Science and Engineering, Hosei University
‡ Faculty of Science and Engineering, Hosei University

3.2 バースト内 tweet からの要約生成

次に，バースト内の状況を要約するために tf-idf を用いる．その際にバーストと判定された範囲内の tweet 全体を 1 文書とみなして解析した．尚，tf-idf 値の高い特徴語を多く含む tweet をバースト毎に選択して時間順に並べたものを要約とした．

4. 評価実験

有効性の検証のために評価実験を行う．

4.1 実験データ

本研究では，Twitter Streaming API を用いて #soccer，#daihyo，#jfa などのハッシュタグを指定して収集した 2014 年 6 月 12 日から 2014 年 7 月 13 日にかけて行われたブラジル開催 FIFA ワールドカップの日本代表戦の tweet データセットを収集した．本研究では試合時間内の 169687 件のデータを対象に解析を行った．

尚，データは 2014 年 6 月 15 日，日本 vs コートジボワール戦:tweet 数 4342 件，ユニークユーザ数:2564，2014 年 6 月 20 日，日本 vs ギリシャ戦:tweet 数 97185 件，ユニークユーザ数:19700，2014 年 6 月 25 日，日本 vs コロンビア戦:tweet 数 68160 件，ユニークユーザ数:9190 となった．

4.2 評価指標

文書要約の評価手法には ROUGE[3]を用いる．ROUGE-N はシステムにおいて要約文書を生成した際に，正解要約と自動生成した要約との間の N グラムの一致する割合を計算するものである．

$$\mathrm{ROUGE(C,R)} = \frac{\sum_{e \in n-gram\,(C)} Count_{clip}(e)}{\sum_{e \in n-gram\,(R)} Count(e)}$$

n-gram(C)は自動生成した要約文書の N グラム，n-gram(R)は正解要約文書の N グラムを示す．Count(e)は，N グラムの出現頻度を数える関数であり，$Count_{clip}$(e)は自動生成した要約における出現頻度 Count(e∈n-gram(C))と正解要約における出現頻度 Count(e∈n-gram(R))の小さいほうを採用する．Lin ら[3]は N を 1 から 4 まで変化させ，検証を行った．その結果 N が 1，2 で最も高い相関があることを示した．久保ら[3]の評価指標同様に Twitter から速報を生成した時間と tweet が紐づいていることから高村らの提案する修正版 ROUGE を使用し，その際 N を 1 とした修正版 ROUGE-1 を用い，時間差を 3 分として名詞，形容詞，動詞のみを用いた．

5.結果

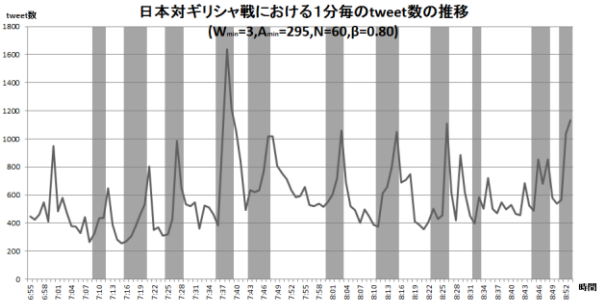

図 3 リアルタイムバースト検出結果

図 3 はリアルタイムバースト検出結果を示している．

表 1 カツラニス退場時のバーストの tf-idf 値

Term	TF-IDF
退場	0.00830
カツラニス	0.00613
アクシデント	0.00538
交代	0.00524
右腕	0.00374
ゲガス	0.00362
イエロー	0.00277
レッド	0.00239
拳	0.00236
寒冷	0.00236

表 1 はギリシャ代表のカツラニス選手がイエローカードを 2 枚出して退場したシーンにおけるバースト内の tf-idf 値の上位 10 件を示している．

表 2ROUGE-1 による評価結果

バースト テキスト	ROUGE-1
1 コートジボワール戦より前向けて、縦パスも入ってますね。いいんじゃないの！？ 前試合より明らかに前を向けてる感触あるよ！	0
2 大迫オフサイド	0.133
3 ギリシャが前からガンガンは来ないので、OBの位置でかなり組み立てられてる。そこから縦パスで大迫が落とし、本田が前へ運ぼうとしたところでファウル。	0.33
4 カツラニスがイエロー2枚目で退場！！これでギリシャが10人になります。日本も吉田選手、長谷部選手がイエロー1枚ずつなので気をつけないと！	0.33
5 前半35分にギリシャのミトログル選手がゲガス選手と交代。アクシデントの模様です。そして37分、カツラニス選手に2枚目のイエロー、退場です！	0.60
6 スタンから外れてしまった香川印のボビッチに入ると真っ先に向かった本田。"大丈夫、絶対蹴ろ"そんな思いがこもってるお人のハグ。	0.33
7 後半から日本代表長谷部誠out→遠藤保仁in	1.00
8 運命の開始45分ｷｯｸｵﾌ！	0.60
9 後半25分過ぎ、ゴール前の混戦で内田が詰めたが外れる。これまた惜しい！ワールドカップギャグの中でも最流行のギャグいいますファール取れないのかよ‥	0.33
10 同時情報を例えトロンディスにイエローカードそしてゴール前でのフリーキックGREE決めろ(●･ω･●)ｸﾞまたフリーキックwwwこれ決めないと業理ぼ、頼む～	1.00
11 アディショナルタイムは4分	0.60

表 2 は日本対ギリシャ戦において生成した要約を ROUGE-1 によって評価した結果を示している．

6.考察

カツラニス選手の退場シーンでは，選手名，退場，イエロー，レッドなどの tf-idf 値が高くなることが確認でき，各バースト内においても，同様に固有の特徴語が確認できた．また．ROUGE-1 による評価結果では正解要約と高い相関を示した．このことから，本手法で，試合全体の状況をよく描写している要約生成が可能であるため，有意味であると考える．

一方で，前半開始の要約に対しては開始以降のバースト内でも投稿がなされており，不向きである．したがって，今後の課題である．

参考文献

1) 蝦名亮平, 中村健二, 小柳滋:リアルタイムバースト検出手法の提案, 日本データベース学会論文誌, Vol.9, No.2, November 2010.

2) 久保光証, 笹野遼平, 高村大地, 奥村学:”良い実況者”に着目した Twitter からのスポーツ速報生成，言語処理学会，第 19 回年次大会，発表論文集.

3) Lin, C.-Y:ROUGE:A Package for Automatic Evaluation of Summaries.Proc, The ACL-04 Workshop“ Text Summarization Branches Out ”,pp.74-81(2004).

4) 坂本翼，廣田雅春，横山昌平, 福田直樹, 石川博:Twitter ストリームのバーストの断続性に着目したキーワード抽出.