D-015

テレビ番組関連 Twitter タイムラインからの代表ツイート選択手法の提案
Tweet Selection Method from Twitter Timeline on TV program

中澤 昌美† 帆足 啓一郎† 小野智弘†
Masami Nakazawa Keiichiro Hoashi Chihiro Ono

1. はじめに

Twitter の普及に伴い，テレビ番組に対する視聴者のツイートを見ながら TV 視聴するという楽しみ方が出てきた．こうした背景から筆者らは，タブレット端末やスマートフォン上に視聴中の番組関連ツイートを番組の進行にあわせて表示する図 1 のようなアプリケーションの開発を進めている．テレビ番組を視聴しながらこのタブレットを流し見することで，例えばスポーツ番組であれば，ネット上でパブリック・ビューイングのような共感体験ができたり，選手の背景知識や苦労話といった新しい番組の見方ができる．

しかし，タブレットやスマートフォンの表示領域には限りがあるため，多数のツイートから表示するツイートを絞り込む必要がある．また，多数派の意見だけではなく，異なる着眼点をもつ意見を抽出するため，ハッシュタグを用いて番組に対するツイートを収集し，各番組で話題となるキーワード（重要語）をもとにツイートをクラスタリングすることで，代表的なツイートを抽出する手法を提案する．

本稿では，バラエティ番組とスポーツ番組を対象に，本手法を適用して抽出した代表ツイートについて調査した結果を報告する．

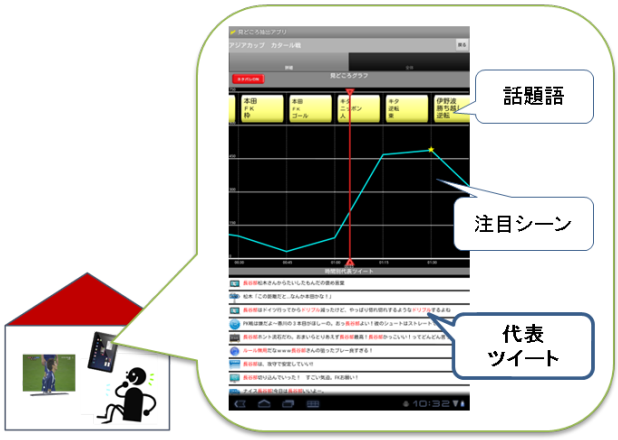

図 1　アプリケーションイメージ図

2. 関連研究とその課題

関連研究として，自動要約技術における複数文書要約が挙げられる．しかし，現在頻繁に行われている研究は，多数の文集合からなる文書の要約が主である．例えば，機械学習による複数文書からの重要文抽出[1]があるが，重要文判定をする際，文章中における重要文の位置情報や長さ，単語の頻度などを評価の尺度としている．ところが，本稿が対象とするマイクロブログは，一人が投稿する文は 1〜数文と短いため，従来手法を適用することはできない．

† ＫＤＤＩ研究所　KDDI R&D Laboratories Inc.

3. 提案手法

本稿の目的は，テレビ番組放送中に投稿された番組に対する代表的なツイートを，録画したテレビ番組を視聴しながら番組の進行に合わせて見ることで，同じような意見に共感したり，異なる視点・着眼点の意見が自動的に得られるアプリケーションを開発することである．そこで，図 2 のフローに示すように，まず，対象となるテレビ番組のツイートを収集する．次に，ツイートの重要語抽出を行うことで着眼点を抽出し，収集したツイートをクラスタリングする．最後に，各クラスタから代表意見を抽出する．以下，それぞれ 3.1，3.2，3.3 で述べる．

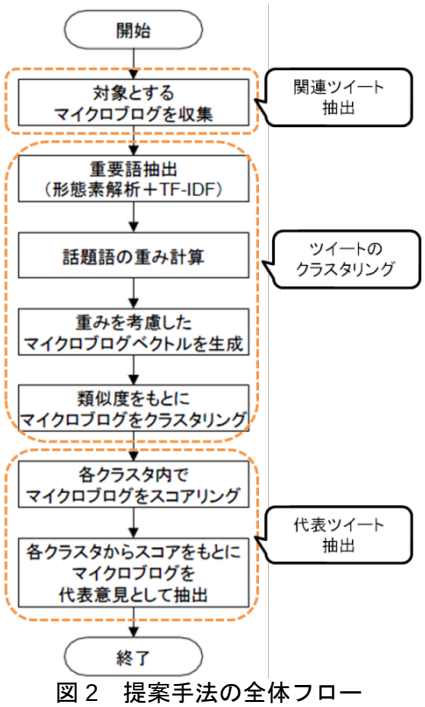

図 2　提案手法の全体フロー

3.1 関連ツイート抽出

関連するツイート抽出では，対象とするテレビ番組の放送中に投稿された番組関連ツイートを収集し，一定時間毎に投稿ツイートを分割する．

具体的には，番組の放送時間帯を対象に，放送局のハッシュタグや，番組で利用されるハッシュタグをキーワードとして検索することで，同一番組に関連するツイートを収集する．次に，ツイートの表示切替時間を設定し，その時間毎にツイートを分割する．デフォルトでは 2 分とし，番組ジャンルやツイート数に応じて変更する．変化の激しいスポーツ番組では，15 秒毎にツイートを切り替える．ドラ

マやバラエティ等の番組では，2 分間あたりの投稿ツイートが 100 件を超えると，1 分毎にツイートを切り替える．このツイート切替設定時間に応じて，番組に関するツイートを一定時間毎に分割する．

3.2 ツイートのクラスタリング

本稿の目的である，様々な意見を代表意見として抽出するために，各区間の投稿ツイートをクラスタリングし，各クラスタから代表的なツイートを抽出する．ここでは，ツイートをクラスタリングする手法について説明する．ツイートをクラスタリングするには，一定区間毎の投稿ツイートからその区間で注目されているキーワード（以下，重要語とする）を抽出する．ツイートに含まれる各重要語の個数により，ベクトルを生成する．その後，生成したベクトルを用いて，各ツイート間の類似度を求め，類似度が高いツイートを同じクラスタに分類していく．以下に，ツイートのクラスタリングの詳細を示す．

まず，重要語抽出を行う．一定区間の投稿ツイートを MeCab によって形態素解析し，抽出された単語に対して TF-IDF 値を求め，上位 N 語を重要語（I_1, I_2,・・・,I_N）として抽出する．重要語の各 TF-IDF 値を，$S(I_1)$,$S(I_2)$,，・・・, $S(I_N)$と表す．ここで DF には，毎日新聞コーパス 10 年分のデータを用いる．

次に，重要語の重み計算を行う．重要語には以下の方法で重みを付ける．

$$W(I_m) = 1 + \frac{S(I_m)}{\sum_{k=1}^{N} S(I_k)}$$

重み行列は，N×Nの行列として，

$$W = \begin{pmatrix} W(T_1) & & 0 \\ & \ddots & \\ 0 & & W(T_N) \end{pmatrix}$$

と表せる．

算出した重みを考慮したベクトルを生成する．各ツイートに対して，重要語が含まれる個数を要素としたベクトルを生成する．例えば，重要語 I_1, I_4 を含むツイートの重みつきベクトルを生成する．ベクトル形式 C_1 で表すと，C_1 = { C_{11}, C_{12}, …, C_{1N}}となる。ただし、C_{1k}は C_1内での重要語 I_kの出現件数とする。さらに，先ほど求めた重み行列をかけて重みつきベクトル C'_1 を生成する．

$$C'_1 = C_1 W$$

$$= \{1,0,0,1,0,0,0,0\} \cdot \begin{pmatrix} W(T_1) & & 0 \\ & \ddots & \\ 0 & & W(T_N) \end{pmatrix}$$

$$= \{W(T_1),0,0,W(T_4),0,0,0,0\}$$

となる．

最後にツイートをクラスタリングする．ここでは，最短距離法による階層型クラスタリングを利用し，予め定めたクラスタ数αになるまで全ツイートをクラスタリングする．以上の操作により，各区間の投稿ツイートをクラスタに分類することができる．

3.3 代表ツイート抽出

代表ツイート抽出では，クラスタ分類したツイートに対しスコアリングを行い，各クラスタからスコアが最も高いツイートを抽出する．

具体的には，各ツイートに含まれる単語の TF-IDF 値の和を算出することで，ツイートのスコアを付け，各クラスタから，スコアが最も高いツイートを抽出することで，各区間からα個の代表ツイートを取り出し，タブレットやスマートフォン等に表示する．

4. 提案手法の分析

本節では，提案手法を用いて抽出したテレビ番組に対する代表ツイートが，共感及び他の着眼点をもつツイートであるか分析する．

あるバラエティ番組（ツイート全 2880 件，放送時間 4 時間）とサッカー中継番組（ツイート全 39815 件，放送時間 2 時間）のツイートを収集し，2 分毎にツイートを分割した．各区間に対し，提案手法によりクラスタリングを行い，代表ツイートを抽出した．重要語は 8 語，クラスタは 8 クラスタに分類した．このデータを用いて抽出した重要語と代表ツイートについての分析を行う．

分析の前に，各番組に対するツイートの傾向を調べると，バラエティ番組では，視聴者は，出演者のトークに対して感じた意見やツッコミを投稿する傾向があり，スポーツ番組では，ゴールやオフサイド等の出来事に応じて，その事実や感想，応援などのツイートを投稿する傾向がある．次に抽出した重要語を調べると，バラエティ番組の重要語 8 件の TF-IDF 値に大きな差がないが，スポーツ番組では差が大きい．バラエティ番組では，重要語 1 位は 8 位の平均 4.8 倍の重要度があるが，スポーツ番組では，平均 12.1 倍の重要度を示す．この理由は，ある選手がゴールを決めると，その選手名を含むツイートが膨大に投稿されるが，バラエティ番組では，一つの話題に集中することは少ないためだと考えられる．抽出したツイートを調べると，バラエティ番組では，抽出される重要語同士に関連する語が多いため，表現は異なるが似通ったコメントが抽出される．異なる着眼点の意見を抽出するという目的のためには，重要語同士の共起頻度を調べる等により，重要語間の関連性を考慮する必要があると考える．反対に，スポーツ番組では，大多数派のツイートに加え，選手の状態や背景知識，目立っていない選手に対するコメントが抽出できたことから，提案手法は一定の有効性が示せたといえる．

5. おわりに

テレビ番組に対する視聴者のツイートをタブレット端末上で見て楽しむためのアプリケーションを開発するため，代表的なツイートを選択する手法を提案した．本稿では，様々な着眼点の意見を代表意見として抽出するために，各区間の投稿ツイートをクラスタリングし，各クラスタから代表的なツイートを選択した．本手法は，バラエティ番組では工夫が必要であるが，スポーツ番組に対しては一定の有効性が示せた．

参考文献

[1] 平尾努 , 賀沢秀人, 磯崎秀樹, 前田英作, 松本裕治, “機械学習による複数文書からの重要文抽出”, 自然言語処理，Vol.10, No.1(2003).