マイクロブログを用いたテレビ番組のリアルタイムなランキング手法
A real-time TV program ranking method using micro-blogs

古宇田　悠輔
Yusuke Kouta
法政大学情報科学部コンピュータ科学科
E-mail: yusuke.kouta.5f@stu.hosei.ac.jp

Abstract

Nowadays more people are posting messages called "tweets" about their actions and thoughts on the Twitter micro-blog service. Since Twitter works in real time, there are many tweets about recent events such as large events, TV programs, and disasters. It is expected that we can quickly get to know about such recent events from Twitter. This is different from the usual TV audience rating because the audience rating does not obtain the opinions of audience. This paper proposes a method that ranks TV programs in real time by collecting tweets about TV programs from Twitter. The method first collects tweets by using basic information about a TV program. Then it creates a keyword list from the collected tweets, and uses it to further collect additional tweets. It uses semantic orientations of words to evaluate whether or not the collected tweets are positive about the TV program. Finally, it ranks TV programs based on the evaluated tweets. This paper also shows the results of experiments on the collect tweets about TV programs.

1. はじめに

近年，マイクロブログ，特に Twitter で，より多くの人々が自身の行動や考えを気軽に発信し，交流を行っている．また，Twitter はリアルタイム性が強く，大規模なイベントやテレビ番組，大きな災害など特定の出来事について多くのツイートが投稿されている．このようなリアルタイムなツイートを用いることによって，特定の出来事の話題性や内容をいち早く察知できると考えられる．

特に多くのツイートがリアルタイムに投稿され，共有されている出来事としてテレビ番組がある．テレビ番組の一般的な指標として視聴率があるが，視聴率には視聴者の意見はなく，情報社会の現代における指標としては弱い．対象となるテレビ番組に関するツイートを利用すれば，視聴率よりも実際の視聴者の意見に近い新たな指標を実現出来る可能性がある．

本論文では Twitter から各テレビ番組に関するツイートを収集し独自のランキングを行う手法を提案する．本手法は最初に特定のジャンルのテレビ番組の情報をテレビ番組のウェブサイトから取得し，その情報をもとにツイートを検索，収集する．リアルタイム性を重視するため，各テレビ番組の放送開始 30 分前からツイートの収集を開始し，番組終了 30 分後まで収集を行う．また，さらに多くのツイートを収集するために，収集したツイートをもとに出現回数の多いキーワードを抽出し，キーワードをもとにツイートを検索，収集する．収集したツイートに関して，対象のテレビ番組についてのツイートか否かを判断し，さらに単語感情極性対応表をもとにテレビ番組について好評価を示しているか否かを判断する．最後に評価結果をもとにテレビ番組のランキングを作成する．本論文では，テレビ番組についてツイートを収集する実験を行った結果を報告する．

2. 関連研究

山本ら [1]はハッシュタグやツイートから作成したキーワード群を用いて対象のテレビ番組についてのツイートをリアルタイムに収集している．しかし，ジャンルごとに 1 つの番組のみの実験しか行っておらず，同一ジャンル内での比較やツイート内容についての考察はなされていない．

汎用連想検索エンジン GETA [2]は連想検索や文書の分類，単語間の類似度計算などの大規模文書の分析を行う．実用例として、文部科学省国立情報科学研究所が提供する Webcat Plus [3]という Web サイトがある。Webcat Plus では文章から図書を連想し検索するシステムである。汎用連検索エンジン GETA は文章の形態素解析を行い、単語ごとの tf-idf による重み付けにより連想検索を行っている。

3. 準備

3.1. tf-idf

本論文の提案手法では，文書の単語ごとの重み付けに tf- idf [4]を使用している．tf (term frequency)は文書内の単語の出現頻度を表し，idf (inverse document frequency)は各文書で単語がどのくらい共通しているかを表すものである．単語数をt，文書数をd，出現回数をn，総文書数をNとし，単語tが出現する文書の出現回数をdf(t)とするとき，tf，idfは以下のように求められる．

$$\mathrm{tf}(t,d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}}$$

$$\mathrm{idf}(t) = \log \frac{N}{df(t)} + 1$$

Supervisor: Prof. Hiroshi Hosobe

tf- idfはtf(t, d)とidf(t)の積である．これによりすべての単語に対して文書における重要度を数値的に示すことができる．

3.2. コサイン類似度

本論文の提案手法では 2 文書間の類似度を測定するためにコサイン類似度 [5]を使用する．コサイン類似度は 2 つのベクトルの類似度を表すために使用される．2 つの文書の特徴をそれぞれベクトルx，yで表すとき，コサイン類似度は以下のように求められる．

$$\cos(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 y_i^2}}$$

コサイン類似度の値が 1 に近いほど 2 つの文書の類似度は高く，0 に近いほど類似度は低い．

4. 提案手法

4.1. 提案手法の概要

提案手法は以下の 4 つの工程からなる．

1. 初めにテレビ番組に関する情報を集め保存する．
2. 次に集めたテレビ番組の情報をもとに対象のテレビ番組に関するツイートを取得する．
3. 集めたツイートをもとに感情の分析を行い，点数化する．
4. 最後に総合的な点数によりテレビ番組のランキングを作成する．

4.2. 番組情報の収集

最初にテレビ番組に関する情報を収集する．テレビ番組の情報はテレビ番組情報サイトから収集する．テレビ番組情報サイトから RSS 形式でウェブページのソースコードを取得し，正規表現を用いて番組情報を取得してテキストファイルに保存する．

4.3. ツイートの収集

本研究の提案手法では，ツイートの収集を 2 段階に分けて行う．

4.3.1. 番組タイトルによるツイートの収集

最初に番組情報をもとにツイートを検索し収集する．初めに対象とするテレビ番組名でツイートを検索する．番組名で検索されたツイートは，対象のテレビ番組についてのツイートである可能性が極めて高い．

4.3.2. キーワードの作成

さらに多くのツイートを取得するために，対象のテレビ番組名から取得したツイートからキーワードを抽出する．テレビ番組に関するツイートには，出演している有名人やキャラクターなどが含まれている可能性が高いため，それらのキーワードを抽出する．

一般に文書から単語を得る手段として形態素解析がある．しかし形態素解析だけでは理想的なキーワードは得られない．例えばキーワードの候補として出演者の名前が挙げられるが，日本人の名前を形態素解析するとほとんどの場合，苗字と名前が別々の単語とされてしまう．

そこで形態素解析を行ったあとに，以下の手順によりキーワードを作成する．最初に各単語の出現回数をカウントする．出現回数の多い単語はキーワードの一部である可能性が高い．出現回数の多い単語を中心に前後の単語を検索していく．初めに中心の単語とその前の単語を連結しキーワード候補とする．生成されたキーワード候補の出現回数をカウントし一定以上の出現回数ならばキーワードとする．次に中心の単語の 2 つ前の単語をキーワード候補に連結し新たにキーワード候補とする．キーワード候補の出現回数が一定未満となった場合，中心の単語の後ろを順に検索し，連結することでキーワード候補の幅を拡張する．

キーワード候補の出現回数が一定未満となったらキーワードとして探索を終了する．

4.3.3. キーワードによるツイートの収集

作成したキーワードを使用して，さらに多くのツイートを収集する．ただし，キーワードを使用して収集したツイートは，本当に対象のテレビ番組に関するツイートなのかという点において，テレビ番組名を使用して収集したツイートよりも正確性が低い．このため，ツイートと番組情報の類似度を用いる．最初に収集した番組情報には番組の出演者やキャラクター名，番組の内容などが記載されているため，対象の番組に関するツイートならばツイートと番組情報の類似度は高いと考えられる．提案手法ではコサイン類似度を使用して計算する．コサイン類似度の計算には，単語の tf-idf からなるベクトルを使用する．ツイートと収集した一週間分の番組情報とのコサイン類似度を計算する。ツイートと対象のテレビ番組とのコサイン類似度が高ければ，そのツイートは対象のテレビ番組についてのツイートと判断する．

4.4. ツイートの評価

4.4.1. 単語感情極性対応表の拡張

ツイートから対象のテレビ番組に関する評価を得るため，単語感情極性対応表 [6]をもとにツイートの評価を行う．しかし，通常の単語感情極性対応表は一般的な文章においてはその評価を十分に発揮するものの，口語表現や Twitter などのネット特有の表現にはあまり対応できていない．そこで単語感情極性対応表にない動詞，形容詞，副詞を，収集したツイートから取り出し，出現回数が一定以上の単語に関して点数を付け，単語感情極性対応表に追加する．これにより単語感情極性対応表を拡張する．

4.4.2. ツイートの感情極性値の算出

拡張した単語感情極性対応表をもとにツイートの感情極性値を算出する．ツイート内に単語感情極性対応表に存在する単語がある場合，単語感情極性対応表の点数をツイートの感情極性値として加算していく．

4.5. ランキングの作成

テレビ番組の評価点をツイートの感情極性値の総和として求め，テレビ番組の評価点の高い順にランキングを作成する．ランキングは任意の 1 週間について作成する．

4.6. 結果提示インタフェース

テレビ番組のランキング結果と各番組について収集したツイートをインタフェースによって提示する．ランキング結果は，指定された任意の 1 週間について提示する．ユーザがランキングの中の番組名をクリックすると，対象の番組について収集したツイートが表示される．

5. 実装

5.1. 番組情報の収集

本研究ではプログラムの実装に Java 言語を使用した．テレビ番組の情報は「G-GUIDE　テレビ王国」 [7]より取得した．RSS 形式で取得したウェブページのソースから正規表現を用いて必要な文字列を取り出す．番組名，番組の詳細，番組の放送日時，番組のジャンルを取り出し，CSV 形式でテキストファイルに書き出す．番組のジャンルはドラマとアニメーションに限定している．番組情報はそれぞれ 1 週間分取得する．取得したての番組名には「[字]」や「木曜ドラマ劇場」などのツイート検索をする際に邪魔になる文字列が含まれているため，正規表現を用いてなるべく番組名を正確に取得している．

5.2. ツイートの収集

Twitter からツイートを収集するために，Twitter API の Java 用ライブラリ Twitter4J [8]を使用した．テレビ番組の情報から番組の放送日時を取得する．テレビ番組の放送時間の 30 分前からツイートの収集が自動的に開始され，番組放送終了 30 分後にツイート収集が終了する．収集したツイートの形態素解析には，形態素解析器 kuromoji [9]を使用した．

5.3. ツイートの評価

収集したツイートの評価のために感情極性対応表を使用する．感情極性対応表は単語ごとにポジティブかネガティブかを点数で段階的に示したものである．最高点が 3 で最低点が−3 である．収集したツイート全てに形態素解析を行い動詞，形容詞，副詞に着目する．着目した単語が単語感情極性対応表にない動詞，形容詞，副詞であれば保存する．保存した各単語についてテレビ番組へのツイートとしてポジティブ・ネガティブを判断できる単語である場合，人手により点数を付け単語感情極性対応表に追加する．単語感情極性対応表の一部を図 1 に示す．

喜ぶ:よろこぶ:動詞:0.999979
褒める:ほめる:動詞:0.999979
めでたい:めでたい:形容詞:0.999645
賢い:かしこい:形容詞:0.999486
善い:いい:形容詞:0.999314
適す:てきす:動詞:0.999295
天晴:あっぱれ:名詞:0.999267
祝う:いわう:動詞:0.999122
功績:こうせき:名詞:0.999104
賞:しょう:名詞:0.998943
嬉しい:うれしい:形容詞:0.998871
喜び:よろこび:名詞:0.998861
才知:さいち:名詞:0.998771
徳:とく:名詞:0.998745
才能:さいのう:名詞:0.998699
素晴らしい:すばらしい:形容詞:0.998617
芳しい:かんばしい:形容詞:0.998578
称える:たたえる:動詞:0.998576
適切:てきせつ:名詞:0.998406

図 1：単語感情極性対応表

5.4. ランキングの作成

ランキングは指定した日付から 1 週間分を作成する．ジャンルはドラマとアニメを対象とする．ランキング作成後の各テレビ番組の評価点は少数第 2 位で四捨五入を行い見やすくする。

5.5. 結果提示インタフェース

結果提示インタフェースは，ドラマとアニメの各ジャンル別にランキングを提示し，番組ごとに収集したツイートを表示する．作成には Java の Swing を使用した．ランキングを提示する期間は， JComboBox で年，月，日付を指定する．指定された日付から 1 週間前の日付を JComboBox の右隣に JLabel で表示する．ランキング結果の出力には JList を使用し，クリックで選択された箇所を視覚的にも分かりやすくしている．また，JList へランキングの結果を出力する際に，順位と単語感情極性対応表により算出された評価点を同時に出力している．ランキングの JList 内の番組タイトルをクリックすると，クリックしたテレビ番組について収集したツイートを表示する．ツイートデータの表示には JTextArea を使用する．ランキング結果の表示に使用した JList とツイートデータの表示に使用した JTextArea にはそれぞれ JScrollPane を適用させ上下左右にスクロールさせることができる．

6. 実験

2014 年 12 月からツイートの収集を開始し，ランキング作成の実験を行った．実験の結果を以下の図 2 に示す．

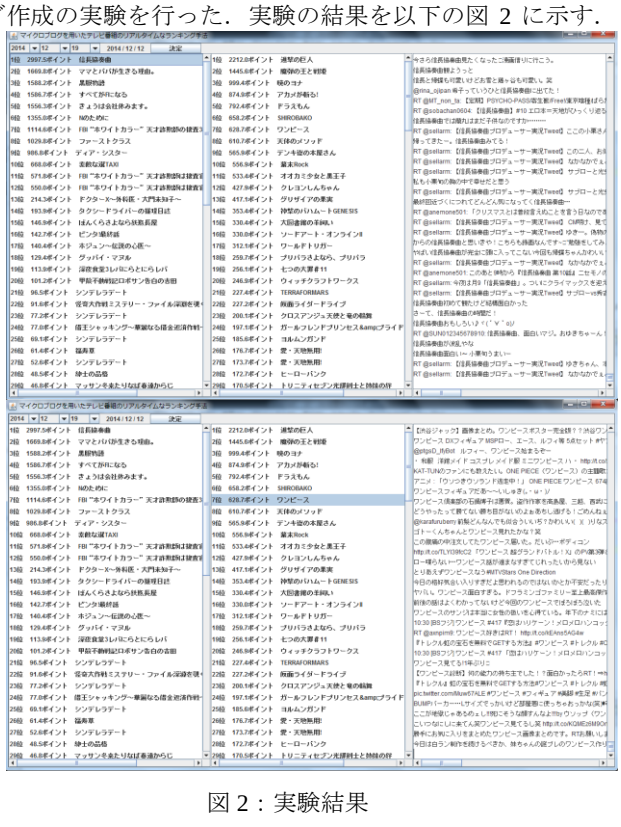

図 2：実験結果

これは 2014 年 12 月 12 日から 2014 年 12 月 19 日の間の各テレビ番組について収集したツイートからランキングを作成したものである．基本的にドラマとアニメは毎週の放送なので，指定した 1 週間のドラマとアニメのランキング結果が出力される．また，図 2 の上側はドラマ部門 1 位の「信長協奏曲」を選択しているため，右のテキストエリアには「信長協奏曲」について収集したツイートがすべて出力されている．一方，図 2 の下側はアニメ部門 7 位の「ワンピース」を選択しているため，右のテキストエリアには「ワンピース」について収集したツイートがすべて出力されている．

また，特定の番組について多くツイートされている時間帯を調べるため番組放送の前後 12 時間と番組放送の 1 時間の計 25 時間について番組についてのツイート数を集計する実験を行った．実験結果を以下の図 3 に示す．

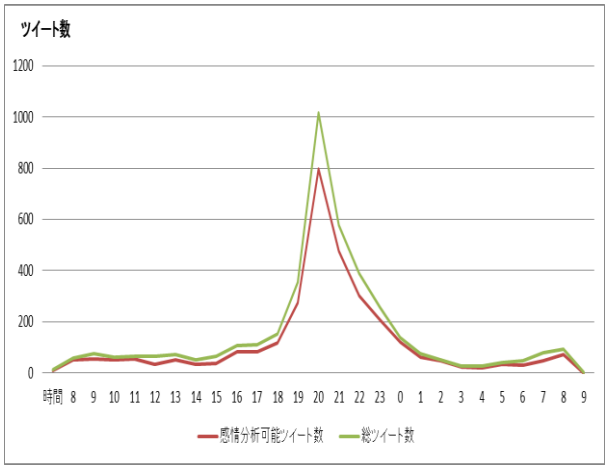

図 3：ツイートの時間帯別収集結果

これは 2014 年 12 月 23 日 21 時から 1 時間放送の「すべてが F になる」について収集したツイート数を時間帯別にグラフにしたものである．縦軸にツイート数を示し，横軸を 1 時間ごとの時間とする．緑色の線は収集したツイート数を示し，赤色の線は，収集したツイートの中で単語感情極性対応表を使用してツイートの感情極性値が算出できたツイート数を示している．番組の放送時間である 21 時から 22 時の間よりも番組放送前の 20 時から 21 時までの間のツイート数がもっとも多いことがわかる．

7. 議論

本論文の提案手法で作成したランキングを視聴率と比較したところ，必ずしも一致する結果にはならなかった．要因は 3 つあると考えられる．

1 つ目の要因はツイート収集が完全ではないことである．例えばアニメ「ワンピース」のツイートの収集結果に，「今日はワンピース着ていこう」などのツイートが収集されていた．このように番組とは無関係なツイートが多数収集されている．また，テレビ番組の公式のツイートやまとめサイトの宣伝ツイートなども収集されており，視聴者の意見ではないツイートが収集されていた．

2 つ目の要因は感情分析の不十分さである．単語感情極性対応表を拡張し充実を図ったが，不十分な点がみられた．例えば「なくはない」という表現は形態素解析後には「ない」が 2 つ存在し，ツイートの投稿者は良い意味で投稿したにもかかわらず，悪い意味となってしまう場合がある．口語表現やネット特有の表現などをさらに対応させる必要がある．

3 つ目の要因は Twitter 利用者の年齢層である．今期のドラマで最も視聴率が良かった番組は，対象年齢層が Twitter 利用年齢層よりも高い傾向があった．また，提案手法で作成したランキングで上位に入った番組は若者を対象とした番組が多い傾向があった．テレビの視聴者全てがインターネットや Twitter などを利用しているわけではない点を考慮しなければならない．

8. おわりに

本研究では，マイクロブログ，特に Twitter を使用して，ツイートを収集，分析し，独自のテレビ番組のランキングを作成した．ツイートの収集の工夫，感情分析の充実，対象とするテレビ番組の年齢層の考慮などが今後の解決すべき課題である．

文　献

[1] 山本祐輔, "テレビ番組に対する意見をもつ Twitter ユーザのリアルタイム検出," 2013.

[2] 増井俊之, "類似ファイル検索," *Unix Magazine/インターフェイスの街角*, no. 12 月, pp. 1-5, 2002.

[3] webcatplus
http://webcatplus.nii.ac.jp/.

[4] takuti, “tf-idf,”
http://blog.takuti.me/2014/01/tf-idf/.

[5] コサイン類似度
http://www.cse.kyotosu.ac.jp/~g0846020/keywords/cosinSimilarity.html.

[6] 感情極性対応表
http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html.

[7] テレビ王国
http://tv.so-net.ne.jp/.

[8] Twitter4j
http://twitter4j.org/ja/.

[9] kuromoji
http://www.atilika.com/ja/products/kuromoji.html.

[10] 高村大也, 乾孝司, 奥村学, "スピンモデルによる単語の感情極性抽出," 2006.

[11] 相澤彰子, "共起に基づく類似性尺度," *オペレーションズ・リサーチ*, vol. 52, no. 11, pp. 706-712, 2007.