2012 年度修士論文

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

提出日： 2013 年 2 月 1 日
指導： 山名早人教授

早稲田大学基幹理工学研究科情報理工学専攻
学籍番号：5111B113-9

山本　祐輔

概　要

近年，テレビ番組を視聴しながら Twitter 等の SNS に意見や感想を投稿する，ソーシャルビューイングと呼ばれる視聴スタイルが盛んに行われている．ソーシャルビューイングにおいて，番組中の人・物・事柄について意見を発信しているユーザの発見は，(1)SNS ユーザにおいて，他人がどのような意見を持っているかを把握できる．(2)テレビ局において，視聴者が番組で着目している人・物・事柄についての意見を得られ，番組制作に反映できるといったメリットがある．本論文では，ソーシャルビューイングを行なっている Twitter ユーザから，意見を持ったユーザを発見する手法を提案する．本手法では，まず，(1)電子番組表及びテレビ番組の字幕放送において表示される字幕テキストから得られる，番組公式の特徴語群，及び，(2)Twitter への投稿からトピックモデルを利用して抽出される，SNS ユーザが生成する番組の特徴語群を得る．そして，番組の放送時間帯に Twitter に投稿されるメッセージに番組の特徴語が含まれているかをチェックすることにより，意見をもった番組実況 Twitter ユーザの検出を試みる ．実験の結果，提案手法は平均して 76%の適合率を保ちながら 68%の再現率で意見投稿を検出することができた．

目　次

第1章　はじめに

近年，Twitter をはじめとするマイクロブログの利用が増加している．マイクロブログとは，情報発信を行うことができるブログの性質とリアルタイムにコミュニケーションがとれるチャットの性質をあわせ持つサービスである．マイクロブログへの投稿は短文で良く，一投稿に必要な時間が短いため，マイクロブログには現在行っていることや考えていることがリアルタイムで投稿される傾向が強い．Twitter はマイクロブログサービスで最も有名なものの 1 つであり，140 文字以内でメッセージを投稿することができる．Twitter は 2006 年にアメリカで開始され，2008 年に日本語化されたサービスであり， 2012 年 12 月の時点で総アカウント数が 5 億を超え[1]，月間アクティブユーザ数は 2 億人を突破している[2].

マイクロブログの流行に伴い，テレビ番組を視聴しながら Twitter 等の SNS に意見や感想を投稿する，ソーシャルビューイングと呼ばれる視聴スタイルが盛んに行われている．Twitter ユーザのうち, 54%がテレビの内容を Twitter に書き込みすることがあり, 30.5%Twitter をきっかけに番組を視聴したとの調査がある[3]．ソーシャルビューイングが行われる背景として，他人の意見を知る，番組の補足情報を提供しあうといった理由のほか，大勢でテレビ番組を見ているような一体感が得られるといった点が挙げられる．最近では，Twitter に投稿されたテレビドラマの感想を，データ放送上に表示させるといった試みや[4]，番組の Facebook ページを作成・運用し，視聴者との双方向のコミュニケーションを行う取り組み[5]など，テレビ番組とソーシャルメディアとの関わりが増加している．また，Twitter のデータを用いて，テレビの視聴率を補完する，視聴率とは異なる影響力指を算出しようという動き[6] [7]や，研究[8][9][10]が見られる．さらに現在，Twitter とテレビ番組の関連性について様々な研究が行われている[11][12][13][14][15][16].

ソーシャルビューイングにおいて，ユーザが Twitter に投稿する内容は多岐にわたるが，大分して(1)感動を表現する投稿，(2)番組中の人・物・事柄についての意見の 2 つに分けることができる．(1)感動を表現する投稿を行う目的としては，大勢でテレビ番組を見ているような一体感が得られ，感動を共有する目的で行わる．一方，(2) 番組中の人・物・事柄についての意見では，自分の意見を発信し，他人と共有する目的で行われる．ここで，(2) 番組中の人・物・事柄についての意見は，ソーシャルビューイングを行なっているユーザにとって他人の意見を知る上で重要な投稿となり，さらに番組理解にも繋がる．また，テレビ局にとっては，視聴率とは異なる視点から番組を評価し，番組の編成に活かすことが可能となる．実際に，Twitter がテレビ番組の制作に活用された例も存在している[17][18].

Twitter には，番組関連ツイートなどをまとめて表示できるハッシュタグという機能がある．ハッシュタグとは，特定のトピックについて投稿する際にメッセージに付加することができるタグ情報である．しかし，ハッシュタグによる，番組中の人・物・事柄について

の意見の抽出には，様々な問題点がある．まず，ハッシュタグはその機能上，タグが付けられた投稿を全て検出するため，感動を表現する投稿と番組中の人・物・事柄についての意見が混在してしまう．また，ハッシュタグを利用するユーザは，Twitter を用いてソーシャルビューイングを行うユーザの 13.8%である[19]．さらに，番組と関係のない投稿にハッシュタグを使用するユーザもいるため，ノイズとなる．

テレビ番組の視聴者がリアルタイムに書き込んだ番組の感想や意見を検出する研究としては，小林らの研究[20]がある．小林らの研究[20]では，番組冒頭時にテレビ番組が始まったことを象徴する投稿が多くなることを利用して，番組開始直後に投稿を行ったユーザに番組を見ているか否かのラベルを付加する．そして，ラベルを付加したユーザが番組放送時間帯に投稿したメッセージ中の単語の出現頻度により，投稿の検出器を作成している．しかし，小林らの研究では，リアルタイムにテレビ番組についての意見を検出することができず，テレビ番組と Twitter のリアルタイム性を活用することが難しい．

そこで，本論文では，ソーシャルビューイングにおいて，リアルタイム性を考慮し，番組中の人・物・事柄についての意見を投稿しているユーザを発見する手法を提案する．本手法では，まず，(1)電子番組表及びテレビ番組の字幕放送において表示される字幕テキストから得られる，番組公式の特徴語群，及び，(2)Twitter への投稿からトピックモデルを利用して抽出される，SNS ユーザが生成する番組の特徴語群を得る．そして，番組の放送時間帯に Twitter に投稿されるメッセージに番組の特徴語が含まれているかをチェックすることにより，意見をもった番組実況 Twitter ユーザの検出を試みる．

本稿では以下の構成をとる．第 2 章では前提知識として Twitter と電子番組表・字幕テキスト及び提案手法で使用する Labeled LDA[21]について説明する．第 3 章で関連研究について述べ，第 4 章で提案手法について説明する．第 5 章で実験と評価を行い，第 6 章でまとめる．

第2章　前提知識

本章では，前提知識として 2. 1 節で Twitter について説明し，2. 2 節で電子番組表・字幕テキストについて述べる．そして，2. 3 節で提案手法に用いる Labeled LDA について説明する．

2. 1　Twitter について

Twitter とは 2006 年にアメリカで開始されたマイクロブログサービスの 1 つである．マイクロブログとは，情報発信を行うことができるブログの性質と，リアルタイムにコミュニケーションがとれるチャットの性質を併せ持つサービスである．マイクロブログへの投稿は短文であるものが多く，一つの投稿に要する時間が短いことが多い．そのため，マイクロブログには現在行っていることや考えていることがリアルタイムで投稿される傾向が強い．

Twitter はマイクロブログで最も有名なサービスの 1 つであり，140 文字以内のメッセージを投稿することができる．Twitter の機能の 1 つで，本研究に関連のあるハッシュタグについて説明する．ハッシュタグとは，ユーザがメッセージを投稿するときに，特定のトピックについての投稿であることを明示するために付加するものであり，トピックを表すキーワードの前に'#'をつけて作成することができる．例えば，NHK で放送されている番組であることを示すために，メッセージに"#nhk"を付けて投稿する． ハッシュタグを使用することにより，特定のトピックについての投稿検索を容易にするというメリットがある．

Twitter のデータは Twitter API により収集することが可能である．API で取得できるデータは JSON 形式で返される．返されるデータからは，投稿内容，投稿したユーザ，投稿された時間などといった様々な要素を取得することができる．本研究に関連のある要素のみを表 1 に記す．本研究で必要な要素は Twitter に投稿されたメッセージの内容，Twitter にメッセージが投稿される時間，およびユーザ名である．本研究では，日本語で投稿されたメッセージを対象とするため投稿内容である"text"にひらがなまたはカタカナが一文字でも含まれるものを採用する．また，Twitter に投稿される時間である"created_at"は協定標準時で表されているため，"utc_offset"を用いて日本標準時に補正する．

また，本研究とは直接関係はないが，次章の関連研究の説明に必要な Twitter の知識を述べておく． あるユーザ A が別のユーザ B の Twitter への投稿を見たいとき，ユーザ A はユーザ B を"フォロー"することにより見ることが可能となる．このとき，A は B の"フォロアー"と呼ばれ，B は A の"フレンド"と呼ばれる．

表 1　API で取得できるデータの要素のうち実験に使用した 要素

API で取得できるデータの要素	要素の意味
“text”	Twitter に投稿されたメッセージの内容
“created_at”	投稿された時間を協定世界時で表したもの
“screen_name”	ユーザ名
“utc_offset”	Twitter にメッセージが投稿された タイムゾーンと協定世界時との差

2.2　電子番組表・字幕テキストについて

電子番組表とは，放送番組表をテレビの画面などに表示するシステムのことである．電子番組表から得られるデータの例を図 1 に示す.

次に，字幕テキストとは，テレビの字幕放送においてテレビ画面に表示されるテキストのことである．字幕テキストは番組内のすべての音声を文字情報として保持している．例えば，番組に登場する人物やナレーションでは,『＞＞おはようございます.』のような字幕が画面に表示される．また，番組中に音楽が流れた場合は『♪～』のような字幕が画面に現れる.

本手法で使用する字幕テキストは，各局の字幕放送で表示されたテキストを，そのテキストが表示された時間とともに記録したものである．使用した字幕テキストの例を図 2 に示す．また，字幕テキストを番組ごとに利用するために，電子番組表から得られる番組の開始時刻と終了時刻を用いる．字幕テキストが表示された情報と電子番組表により，番組ごとに字幕テキストを切り出すことができる.

字幕テキストを番組単位で分けることにより，字幕テキストから番組の特徴となる語を取得することができると考えられる．また，電子番組表にも番組のタイトル(図 1 の title)や概要(図 1 の description)から番組の特徴語を得られる．特に，番組のタイトルは番組を視聴しているユーザを特定する手がかりとなりやすい．しかし，字幕テキストはドラマにおけるキャスト等の番組中に音声として現れない情報が欠けているため，番組に関連する全ての情報を網羅しているわけではない.

iepg	station	start	end	title	genre	subgenre	description
101024201203221605	NHK総合	1332399900	1332406800	[二][字]大相撲春場所　−十二日目−「優勝争い占う戦い」	16	8	「優勝争い占う戦い　白鵬・琴欧洲　把瑠都・琴奨菊　鶴竜・翔天狼」（4:10)「幕内取組」【解説…
101040201203221653	日本テレビ	1332402780	1332410400	news　every.[字]	69	1	スカイツリー予約開始▽ウナギ回遊"異変"▽全国地価公表1位は▽日本全国名物調べ隊発見"ご当地珍料理"…
101064201203221653	テレビ朝日	1332402780	1332410400	スーパーJチャンネル	69	1	スカイツリー抽選開始"殺到"に不安の声も▽目玉は"1円福袋"ブランド市に客殺到!!先着100人が大争…
101048201203221653	TBSテレビ	1332402780	1332410400	Nスタ[字]	69	1	ゲーム会社社長は12歳▽分裂か…消費税法案▽スカイツリー入場券予約開始入手方法は？▽カバはバンでサル…

図 1　電子番組表から得られるデータの例

id	time	serviceid	caption
23743966	2013-01-13 21:00:16	1024	ウオ
23743969	2013-01-13 21:00:16	1024	〜
23743972	2013-01-13 21:00:16	1024	ッ
23743974	2013-01-13 21:00:16	1024	！
23743976	2013-01-13 21:00:16	1024	あ
23743978	2013-01-13 21:00:16	1024	
23743980	2013-01-13 21:00:16	1024	来た
23743982	2013-01-13 21:00:17	1024	。
23743993	2013-01-13 21:00:23	1024	世界最大のイカ
23743994	2013-01-13 21:00:23	1024	ダイオウイカ。
23744016	2013-01-13 21:00:27	1024	史上
23744017	2013-01-13 21:00:27	1024	
23744018	2013-01-13 21:00:27	1024	初めて
23744019	2013-01-13 21:00:27	1024	深海での撮影に成功しました
23744020	2013-01-13 21:00:27	1024	。
23744026	2013-01-13 21:00:32	1024	ダイオウイカ
23744027	2013-01-13 21:00:32	1024	は
23744028	2013-01-13 21:00:32	1024	1，000
23744029	2013-01-13 21:00:32	1024	年もの間→
23744030	2013-01-13 21:00:35	1024	伝説の怪物として
23744031	2013-01-13 21:00:35	1024	恐れられてきました
23744032	2013-01-13 21:00:35	1024	。
23744049	2013-01-13 21:00:41	1024	巨大な体で
23744050	2013-01-13 21:00:41	1024	船を丸ごと引きずり込む
23744051	2013-01-13 21:00:41	1024	。
23744063	2013-01-13 21:00:45	1024	船乗りたちを
23744064	2013-01-13 21:00:45	1024	震え上がらせてきたのです
23744065	2013-01-13 21:00:45	1024	。

図 2　字幕テキストの例(NHK 2013 年 1 月 13 日分の一部)

2. 3 Labeled LDA

本節では Labeled LDA について説明する．Labeled LDA とは，文書中のトピックを示すラベルが付与されている文書集合の生成過程をモデル化する，確率的トピックモデルの 1 つである．Labeled LDA では，文書はトピックの出現確率を表す多項分布として表現され，トピックは単語の出現確率を表す多項分布として表現される．この時，トピックは文書集合に付与されたラベルと一対一に対応する．つまり，各文書中の単語は，その文書に付与されたラベルが表すトピックから生成される．

文書集合中の文書dを，単語リスト$\boldsymbol{w}^{(d)} = \{w_1, \dots, w_{N_d}\}$及びトピックの有無を表すリスト$\boldsymbol{\Lambda}^{(d)} = \{l_1, \dots, l_K\}$により表せるとする．ただし，$d \in D$，$w_i \in V$，$l_k \in \{0,1\}$である．ここで，$N_d$は文書$d$の単語数，$D$は文書集合，$V$は文書集合中の語彙集合，$K$は文書集合に付与されているユニークなトピック集合である．このとき，文書の生成過程のアルゴリズムは以下のようになる．

(1) 各トピック$k \in K$について：
  (a) ディリクレ分布$Dir(\boldsymbol{\eta})$に従って単語の出現確率を表す多項分布$\boldsymbol{\beta}_k$を生成
    $\boldsymbol{\beta}_k \sim Dir(\boldsymbol{\eta})$
(2) 各文書$d \in D$について：
  (a) 各トピック$k \in K$について：
    i. ベルヌーイ分布$Bernoulli(\Phi_k)$に従い，ラベルの有無$\Lambda_{\mathrm{k}}^{(d)}$を決定
      $\Lambda_{\mathrm{k}}^{(d)} \sim Bernoulli(\Phi_k)$
  (b) ラベルに対応するディリクレ分布のパラメータ$\boldsymbol{\alpha}^{(d)}$を生成
    $\boldsymbol{\alpha}^{(d)} = L^{(d)} \times \boldsymbol{\alpha}$
  (c) ディリクレ分布$Dir(\boldsymbol{\alpha})$に従って単語の出現確率を表す多項分布$Mult\boldsymbol{\theta}^{(d)}$を生成
    $\boldsymbol{\theta}^{(d)} \sim Dir(\boldsymbol{\alpha})$
  (d) $i \in \{1, \dots N_d\}$について：
    i. 多項分布$Mult(\boldsymbol{\theta}^{(d)})$に従いトピック$z_i$を生成
      $z_i \sim Mult(\boldsymbol{\theta}^{(d)})$
    ii. 多項分布$Mult(\boldsymbol{\beta}_{z_i})$に従い単語$\mathrm{w}_i$を生成
      $\mathrm{w}_i \sim Mult(\boldsymbol{\beta}_{z_i})$

ここで，生成過程中の$L^{(d)}$について説明する．まず，文書dに付与されているラベルを表すベクトル$\boldsymbol{\lambda}^{(d)} = \{k | \Lambda_{\mathrm{k}}^{(d)} = 1\}$を定義する．これを用いて，文書固有の射影行列$L^{(d)}$を以下の式(1)で定義できる．ただし，$i \in \{1, \dots, |\boldsymbol{\lambda}^{(d)}|\}$，$j \in \{1, \dots, \mathrm{K}\}$である．

$$L_{ij}^{(d)} = \begin{cases} 1 \text{ if } \lambda_i^{(d)} = j \\ 0 \text{ otherwize} \end{cases} \quad \textbf{(1)}$$

例えば$K = 4$，文書dに付けられているラベルの有無を表すリストを$\boldsymbol{\Lambda}^{(d)} = (0,1,1,0)$とすると，$\boldsymbol{\lambda}^{(d)} = \{2,3\}$となり，$L^{(d)}$は$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$となる．

つまり，生成過程のアルゴリズムにおけるステップ(2)-(b)では，射影行列$L^{(d)}$を用いて，ディリクレ分布のパラメータベクトル$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\boldsymbol{T}}$を低次元のベクトル$\boldsymbol{\alpha}^{(d)}$に射影している．$\boldsymbol{\alpha}^{(d)}$を以下の式(2)に示す．

$$\boldsymbol{\alpha}^{(d)} = L^{(d)} \times \boldsymbol{\alpha} = (\alpha_{\lambda_1^{(d)}}, \ldots, \alpha_{\lambda_{|\boldsymbol{\lambda}^{(d)}|}^{(d)}})^{\boldsymbol{T}} \quad \textbf{(2)}$$

Labeled LDA[]において，パラメータ$\boldsymbol{\beta}_k$，$\boldsymbol{\theta}^{(d)}$の推定にはCollapsed Gibbs Sampling[]が使用されている．本稿においても，Collapsed Gibbs Samplingにより，パラメータ$\boldsymbol{\beta}_k$，$\boldsymbol{\theta}^{(d)}$の推定を行う．Labeled LDAにおけるCollapsed Gibbs Samplingの更新式を以下の式(3)に示す．

$$P(z_i = j | \boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n_{-i,j}^{w_i} + \eta}{n_{-i,j}^{(\cdot)} + V\eta} \times \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha} \quad \textbf{(3)}$$

ただし，$\boldsymbol{z}_{-i}$は$\boldsymbol{z}$からz_iを除いたもの，$n_{-i,j}^{w_i}$は位置i以外でトピックjから単語w_iが生成された回数，$n_{-i,j}^{(d)}$は位置i以外で文書dにおいてトピックjが現れた回数，$n_{-i,j}^{(\cdot)}$は位置i以外でトピックjがコーパス全体で現れた回数，$n_{-i,\cdot}^{(d)}$は位置i以外で文書dに含まれる単語数である．サンプリングによって得られたサンプルから，書くトピックの単語分布$\boldsymbol{\beta}_k$と各文書のトピック分布$\boldsymbol{\theta}^{(d)}$を推定する．推定されるパラメータ$\widehat{\boldsymbol{\beta}}_k$，$\widehat{\boldsymbol{\theta}}^{(d)}$はそれぞれ以下の式(4)，式(5)により求められる．

$$\hat{\beta}_{j,w} = \frac{n_j^w + \eta}{n_j^{(\cdot)} + V\eta} \quad \textbf{(4)}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{(d)} + K\alpha} \quad \textbf{(5)}$$

第3章　関連研究

本章では，本研究と関連している研究について述べる．3. 1 節で Twitter とテレビ番組との関連性について調査した研究について説明し，3. 2 節でテレビ番組についてリアルタイムに書き込まれるコメントを対象とした研究について述べる．そして，3. 3 節では Twitter からテレビ番組に言及している投稿を検出する研究という，本研究と最も関連する研究について述べる．

3. 1　Twitter とテレビ番組との関連性についての研究

Twitter にはテレビやニュースサイトで報じられた情報に対する投稿が多数行われている．そのため，Twitter と他のメディアとの関連性が研究の対象となっている．本節では，Twitter と他のメディアとして代表的なテレビ番組との関連性についての研究を紹介する．まず，3. 1. 1 項で Shamma らが行ったライブメディアイベントに対する Twitter の使われ方の研究[11]について述べる．3. 1. 2 項では，秋岡らが行った日本における Twitter ユーザの特徴および他のメディアが Twitter に与える影響についての研究[12]について説明し，0 項では，澤井らが行った Twitter を用いたテレビ番組の推薦についての研究[13]を紹介する．0 項では，加藤らが行った Twitter を用いたテレビ番組で注目される話題を追跡する研究[14]について述べ，3. 1. 5 項では Ariyasu らが行ったテレビ番組に対する投稿を解析するシステムを作成した研究[15]について説明し，3. 1. 6 項では，Hu らが行ったライブメディアイベントに対する Twitter への投稿を分類する研究[16]を紹介する. そして, 0 項で Wakamiya らが行った Twitter の投稿を用いてテレビ番組を評価する研究[8][9][10]について述べる．

3. 1. 1 Shamma らの研究[11]

2009 年に Yahoo! Research の Shamma らは，2008 年のアメリカ大統領選挙において行われた候補者同士のディベートの生放送を題材として次の 2 つの調査を行った．ここで，調査に使用した Twitter のデータは，ハッシュタグをもとに収集している．

1. Twitter に投稿されるメッセージからディベートを意味単位に分割する．つまり，ディベートするトピックの変化を予測できるかどうかに関する調査である．
2. ユーザ同士のメッセージのやりとりを追跡することにより，重要なユーザを発見する調査する．

まず，ディベートするトピックの変化を予測する調査について説明する．まず，Shamma らは前提として映像システムとチャットに関する文献[22]を利用した．この文献[22]では，映像コンテンツについて人々が最も盛んに議論し合うのは，ビデオが終わった時だと報告している．そのため，Shamma らは Twitter への単位時間当たりの投稿数にスパイクが発生している時が，トピックが変化した場所だと定義した．そして，次に示す手法により，トピックの変化点を捉えた．

step1. ディベートのハッシュタグが付与されている投稿数を毎分計測する．

step2. 計測時点及びの計測時点の前後 1 分間の計 3 分間の投稿数の平均を取り，グラフにプロットする．

step3. 計測時点及び計測時点の前後 10 分間の計 21 分間の投稿数から平均μと標準偏差σを算出する．

step4. step2.でプロットした値がμ±σの範囲外にある極値をトピックの変化点とする．

Shamma らの実験により推定したトピックの変化点を図 3 に表す．実験の結果，前後 1 分の誤差でトピックの変化を捉えることに成功したと報告している．

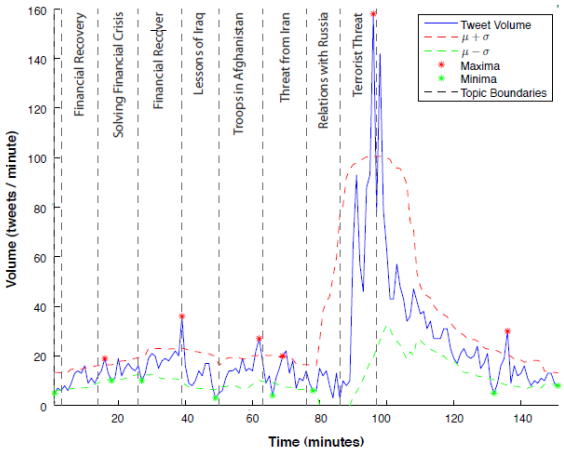

図 3　推定したトピックの変化点([11]の Figure 3 より引用)

次に，ユーザ同士のメッセージのやり取りを追跡することにより重要なユーザを発見する調査について説明する．Shamma らは次に示す手法により重要なユーザを発見している．

step1. ディベートのハッシュタグが付与されている投稿からメンションツイートを抽出する．

step2. メンションツイートの送信ユーザから受信ユーザに有向リンクを張り，グラフ構造を作成する．

step3. step2.で作成したグラフに固有ベクトル中心性を適用する．

step4. 固有ベクトル中心性が高いユーザを重要ユーザとする．

step2.で作成されるグラフを図 4 に表す． 実験の結果，ディベートの参加者やモデレータが重要なユーザとなったと報告している．

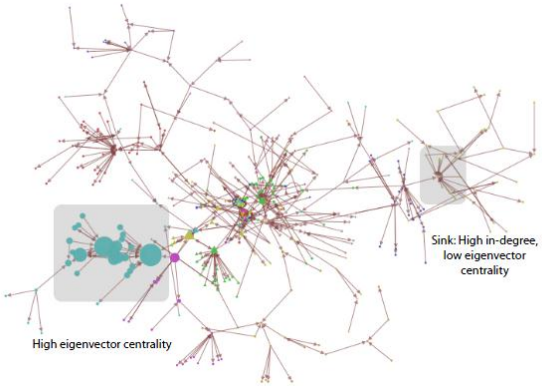

図 4　メンションツイートにより作成されるグラフ([11]の Figure 5 より引用)

3. 1. 2　秋岡らの研究[12]

2010 年に早稲田大学の秋岡らは日本における Twitter ユーザの特徴および出版物やテレビ等のメディアが Twitter のコミュニティに与える影響について研究している．

まず秋岡らが，日本における Twitter ユーザの特徴について行った研究の説明を行う．秋岡らはプロフィールが日本語で書かれている 50,000 以上のユーザの情報を収集した．日本人の Twitter ユーザは 5,000,000 人と言われているため，約 1%の日本ユーザの情報を取得している．取得した情報を用いて以下に示す 3 つの研究を行っている．

1. Twitter ユーザが持つ統計情報の傾向を把握する．(フレンド数の傾向，フォロワー数の傾向，及びフォロワー数に対するフレンド数の傾向)
2. フォロアー数が多い Twitter ユーザのランキング，及びフォロー関係をユーザ間のリンクとみなし，PageRank アルゴリズムによる Twitter ユーザのランキングを行う．
3. 出版物やテレビ等のメディアが Twitter のコミュニティに与える影響について調査する．

まず，Twitter ユーザが持つ統計情報の傾向について説明する．まず，Twitter ユーザのフレンド数の傾向(図 5)と，フォロワー数の傾向(図 6)は，負の相関関係を示した．つまり，フレンド数やフォロワー数が多くなるほど，対応する数のフレンドやフォロワーを持つユーザ数は少なくなっている．また，フォロアー数に対するフレンド数は正の相関関係を示した(図 7)．つまり，ユーザが持つフォロアー数が多い場合，そのユーザが持つフレンド数は比例して多くいることを表している．

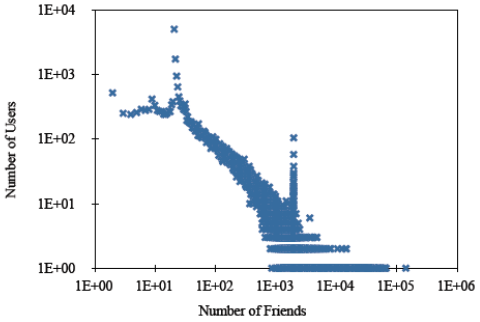

図 5　フレンド数の傾向([12]の Figure 1 より引用)

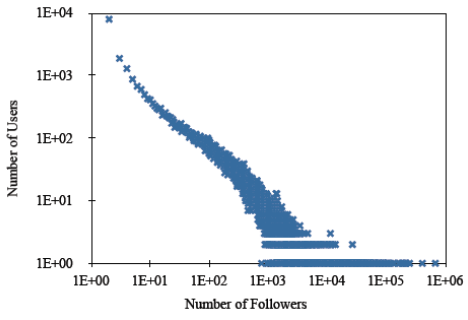

図 6　フォロワー数の傾向([12]の Figure 2 より引用)

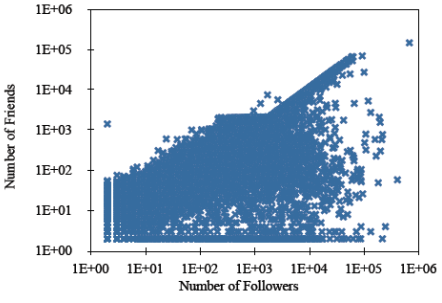

図 7 フォロワー数に対するフレンド数の傾向([12]の Figure 3 より引用)

次に，秋岡らが行ったユーザのランキング結果についての説明を行う．フォロアー数が多いユーザによるランキングと PageRank によるランキングでは，ともに有名な企業の CEO や政治家が上位にランキングされた．しかし，秋岡らは，両者のランキング結果には次に示す 2 つの違いがあると報告している．

1. PageRank によるランキング結果では，アルゴリズムの特徴により，フォロアー数の多いユーザからフォローされているユーザが上位にランキングされる．
2. フォロアー数が多くいるユーザで，かつフォロアー数とフレンド数が同程度の場合，PageRank では上位にランキングされている．

最後に，出版物やテレビ等のメディアが Twitter のコミュニティに与える影響について行われた調査について説明する．秋岡らは，Twitter に関する出版物やテレビ番組が Twitter の新規ユーザ数に影響を与えたかどうかについても調査している．調査の結果，新規ユーザ数の増加には以下に示す 3 つの特徴があると結論づけられている．

1. Twitter に関する出版物は，インターネットのヘビーユーザを Twitter に参加させている．
2. Twitter が日本語された直後に新規ユーザが急増する．
3. テレビ番組で Twitter が取り上げられると，その後 2,3 日は新規ユーザが増加する．

3. 1. 3　澤井らの研究[13]

2010 年に NHK 放送技術研究所の澤井らは, Twitter に投稿されるメッセージを利用して, 協調フィルタリングによりテレビ番組を推薦する研究を行った．澤井らは，番組の推薦手法を提案した．提案手法では図 8 に示す状態を前提としている．

まず，澤井らが提案した 1 つ目の手法について説明を行う．1 つ目の提案手法をまとめると以下の 5 つの step からなる．

step1. 番組推薦を行う基準となる代表的なユーザを抽出する．

step2. 番組を推薦する対象であるユーザのプロファイルを作成する．

step3. step2 で推薦対象ユーザのプロファイル作成に利用されるユーザが視聴している番組名を推定する．

step4. 推薦対象ユーザのプロファイルの特徴ベクトルを作成する．

step5. 推薦候補番組を表現する特徴ベクトルを作成する．そして，推薦対象ユーザのプロファイルの特徴ベクトルと類似する特徴ベクトルを持つ番組を推薦する．

次に，澤井らが行った実験について述べる．実験では，提案手法により，日常的に SNS を利用しないユーザへの NHK 総合で放送される番組の推薦が行われた．実験の結果，視聴率の高い番組だけでなく，視聴率が低くても Twitter で注目されている番組が推薦された．また，NHK 総合では，ニュース番組の割合が高いが，さまざまなジャンルの番組が推薦されたと報告している．

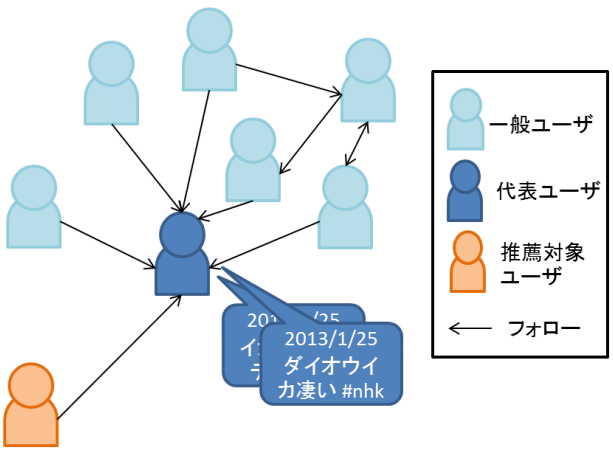

図 8　番組推薦における前提の状態([13]の図 2 を参考に作成)

3. 1. 4　加藤らの研究[14]

2010 年に早稲田大学の加藤らは，Twitter に投稿されるメッセージから出現頻度が急に高くなった固有名詞を抽出し，テレビ番組で放送される話題を追跡する研究を行った．

まず，出現頻度が急に高くなった固有名詞を抽出する方法から説明する．以下，出現頻度が急に高くなった固有名詞を注目語と呼ぶ．加藤らは Twitter に投稿されるメッセージから出現頻度が急増する単語(急上昇語)を取り出し，以下に示す手法により注目語を取得している．

step1. 急上昇語を含んでいる投稿を抽出する．

step2. 投稿を急上昇語より前と後に分割する．

step3. 分割後の文章を形態素解析し，各単語に「急上昇語から n 個目の形態素」というラベルを付加する．

step4. 急上昇語に近い形態素から順に，各ラベルで最も頻出する単語が，急上昇語を含んでいる投稿に含まれている割合を求める．

step5. step4 で求めた割合が閾値以下となるまで探索を行う．

次に，取得した注目語を用いて，テレビで放送された話題を追跡する方法について説明する．

step1. 提案手法により話題となっている注目語を抽出する．

step2. step1.で得られた注目語を形態素解析辞書に追加する．

step3. 注目語を追加した辞書を用いて，字幕テキストを形態素解析する．

step4. step3.の結果からテレビの番組に出現する固有名詞を得る．

Twitter から取得した注目語を追加した形態素解析辞書による字幕テキストの解析は，従来の形態素解析辞書と比較し 7%多くの固有名詞を取得している．この結果より，加藤らは Twitter からの注目語の抽出は，テレビで放送されている話題における注目語の抽出に有効であり，また，Twitter で注目されている話題とテレビで放送されている話題には相関があると結論付けている．

3. 1. 5　Ariyasu らの研究[15]

NHK 放送技術研究所の Ariyasu らは，　Intelligence Circulation System(以下 ICS とする)を実現するために，テレビ番組についての投稿を解析するアルゴリズムについて論じている．ICS とは Twitter への投稿を利用し，番組のトレンドグラフ(図 x)や投稿を基にしたダイジェストアニメーションの作成，ユーザへの番組推薦(図 x)を行うシステムのことである．なお，番組に関連する投稿は，番組のハッシュタグを利用して収集している．

Ariyasu らは ICS を実現する上で，Twitter への投稿を解析する 3 つのアルゴリズムを提案している．

1. 番組に関連する投稿からその投稿が示すトピックを検出する．検出方法は，投稿文に EPG および字幕テキスト中の語句が含まれている場合その語句をトピックとし，含まれていない場合はトピックが定まっている他の投稿との類似度を計算し，最も類似している投稿のトピックを使用している．
2. 番組に関連する投稿に対し，単語辞書ベースの感情分析を行い，投稿が positive か negative か neutral かを判断する．
3. Twitter への投稿に要する時間が，投稿内容とそれに対応する放送時間との誤差となっているため，時間の補正を行う．

実験の結果，70%の確率で番組に関連する投稿からトピックを検出することに成功し，感情分析では精度 85%，再現率 74%になったと報告している．

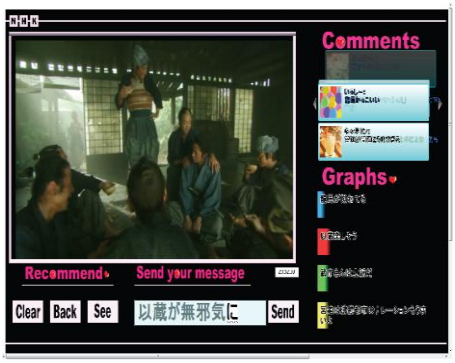

図 9　ICS におけるトレンドグラフ([15]の図 5 より引用)

図 10　ICS のおけるユーザへの番組推薦([15]の図 7 より引用)

3. 1. 6　Hu らの研究[16]

2012 年にアリゾナ州立大学の Hu らはアメリカで放送された大統領スピーチおよびディベートの 2 つのイベントにおいて，Twitter に投稿されたメッセージを解析し，イベント中の特定の内容についての投稿(以下 *episodic* な投稿とする)なのかイベント全体についての投稿(以下 *steady* な投稿とする)なのかを分類する研究を行なっている．

Hu らは，対象としたイベントに関連する投稿を，イベントのハッシュタグを元に収集した．この際，イベント開催時間帯およびのその前後 5 時間に渡り投稿を収集している．これは，イベントの前後においても，イベントに関連する投稿がなされるためである．

その結果として Hu らは，以下の 3 つの結果を報告している．イベントにおける *episodic* な投稿の推移を図 11 と図 12 に示す．

1. *episodic* な投稿はイベントの開催時間帯に多く投稿されるが，イベントの前後ではそれほど投稿されない．図 11 と図 12 より，イベント開催中には *episodic* な投稿の割合が約半分を占めるが，イベントの前後はそれぞれ 35%と 38%しかない．
2. Twitter への投稿数及び *episodic* な投稿数から，ユーザがどれだけイベントに関心を持っているかわかる．
3. イベントの開催時間帯に渡り，あらゆる *episodic* な投稿が行われる．つまり，ディベートにおいて，経済についてディベートされる前から経済についての *episodic* な投稿があり，また，経済についてディベートされた後にも経済についての *episodic* な投稿がある．

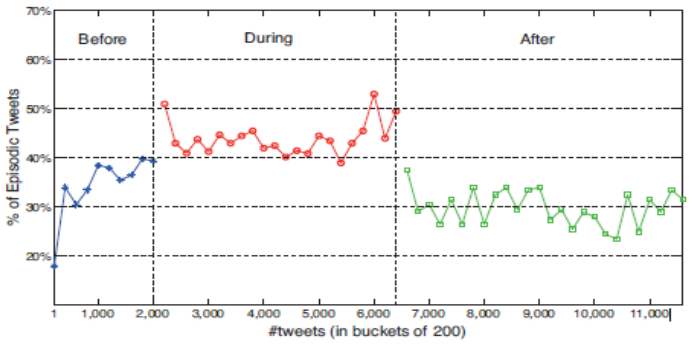

図 11　大統領選におけるスピーチでの *episodic* な投稿の割合([16]の Figure4 より引用)

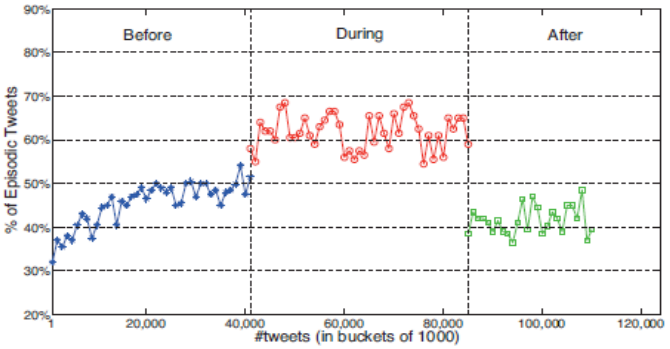

図 12　大統領選におけるディベートでの *episodic* な投稿の割合([16]の Figure4 より引用)

3. 1. 7　Wakamiya らの研究[8][9][10]

兵庫県立大学の Wakamiya らは，Twitter の投稿を用いて，視聴率とは異なるテレビ番組評価手法を提案している．

Wakamiya らが，テレビ番組を評価する手順は以下に示す 3 つの step である．また，図 13 にシステムの流れを示す．

step1. 位置情報が付けられた投稿を収集する．

step2. 以下の 3 つの指標により Twitter への投稿が，テレビに関連するものであるか判断している．

i. 投稿に含まれる名詞と番組タイトルに含まれる名詞の類似度

ii. 投稿を行った場所とテレビ局の場所との距離

iii. 投稿を行った時間とテレビ番組の放送時間

step3. テレビに関連している投稿数$\#tweet$及びそれらの投稿を行ったユーザ数$\#user$を用い，以下の式**(6)**によりテレビ番組e_jの人気度$popularity(e_j)$を評価している．

$$popularity(e_j) = \sqrt{\#tweet} \times \#user \qquad \textbf{(6)}$$

実験の結果，人気度が高い番組のジャンルは，トークショー，アニメ，ドラマであると報告している．

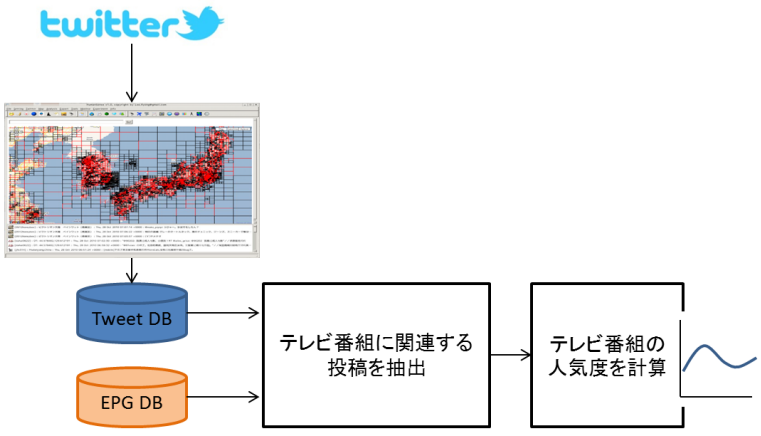

図 13　テレビ番組の評価システムの流れ([10]の Figure9 を参考に作成)

3. 1. 8　Twitter とテレビ番組との関連性について調査した研究のまとめ

本節では，Twitter とテレビ番組との関連性についての研究を紹介した．本節で紹介した研究を表 2 にまとめる．表 2 中の『番組に言及している投稿の判定方法』の欄に斜線が入っている研究は，Twitter とテレビ番組についての研究であるが，番組に言及している投稿を利用していない研究である．

表 2 Twitter とテレビ番組との関連性についての研究のまとめ

研究	研究内容	番組に言及している投稿の判定方法
Shamma らの研究[11]	テレビで生放送されるイベントについて，Twitter へ投稿されるメッセージを解析し，イベント内容の理解を促す研究	テレビ局とテレビ番組のハッシュタグ
秋岡らの研究[12]	日本における Twitter ユーザの特徴および出版物やテレビ等のメディアが Twitter のコミュニティに与える影響についての研究	
澤井らの研究[13]	Twitter を利用して，テレビ番組を推薦する研究	テレビ局のハッシュタグおよび人手により判断した番組のキーワード
加藤らの研究[14]	Twitter から注目されている語を抽出し，テレビ番組で放送されている話題を追跡する研究	
Ariyasu らの研究[15]	テレビ番組に対する投稿を解析するシステムの作成	テレビ番組のハッシュタグ
Hu らの研究[16]	ライブメディアイベントに対する Twitter への投稿を分類する研究	テレビ番組のハッシュタグ
Wakamiya らの研究[8][9][10]	Twitter の投稿を用いてテレビ番組を評価する研究	投稿の位置情報と電子番組表から得られる番組タイトル名

3. 2　テレビ番組に関してリアルタイムに書き込まれるコメントを対象とした研究

テレビ番組に関する内容がリアルタイムに書き込まれる Web ページとして，掲示板サイトが有名である．掲示板サイトへの書き込みは有用な情報を抽出する研究に利用されている．本節では，テレビ番組に関してリアルタイムに書き込まれるコメントを対象とした研究として，3. 2. 1 項で宮森らが行った掲示板サイトからテレビ番組のビューを生成する研究[23]を紹介し，3. 2. 2 項で上原らが行った掲示板サイトから番組で注目されている人物や事柄をグラフ化する研究[24]を紹介する．

3. 2. 1　宮森らの研究[23]

2005 年に情報通信研究機構の宮森らは，番組の感想や意見がリアルタイムに書き込まれる掲示板への書き込みを統計処理・認識処理することにより，番組の盛り上がり場面や，特定の視聴者，例えば自分と思考が類似しているユーザが興味を示している場面等を抽出し，シーン探索やダイジェスト視聴において視聴者の視点を取り入れられるようにする研究を行った．

宮森ら提案するシーンのインデキシング処理は以下に示す step により行われる．また，システムの流れを図 14 に示す．

step1. 記録されたコメントのデータをパースすることにより書き込み時刻，書き込んだユーザの ID，書き込みの内容を得る．

step2. 次に，書き込み時刻が書き込み内容のシーンより遅れることの補正を行う．

step3. 単位時間当たりのコメント数より反響の大きさを計算する．また，感情を表す ASCII アートやフレーズから盛り上がりや落胆の大きさを計算する．

インデキシングされたシーンを利用することにより，例えば単位時間あたりの書き込み数によりソートを行い，ユーザの反響が大きかったシーンを視聴する際に役立てることができる(図 15)．また，コメントを書き込んだユーザ ID の情報から，特定の視聴者の視点によりシーンをランキングし，自分と類似した嗜好を持っているユーザをはじめ，反対の価値観を持つユーザの視点からシーンを視聴することも可能となる．

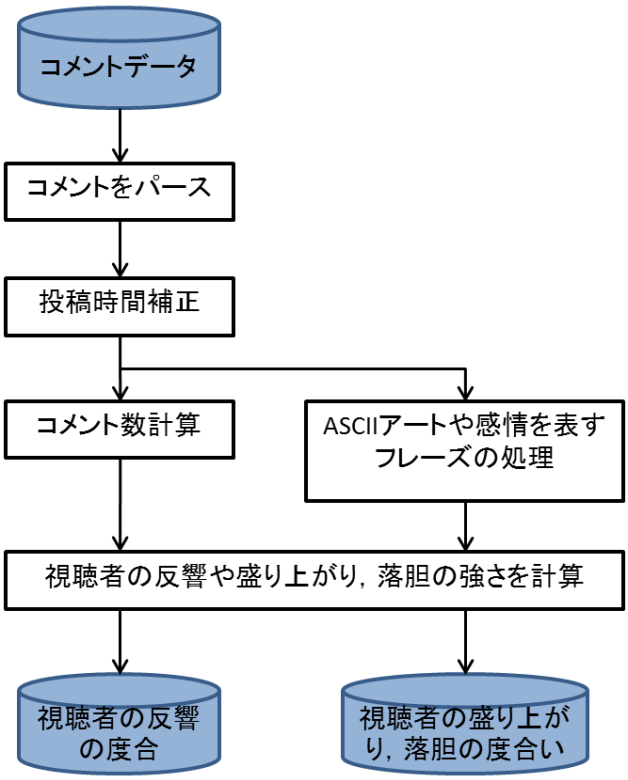

図 14 シーンのインデキシング処理の流れ([23]の図 4 を参考に作成)

図 15 ユーザの反響によりランキングされたシーン([23]の図 11 より引用)

3. 2. 2　上原らの研究[24]

2004 年に筑波大学の上原らは，掲示板サイトに書き込まれるテレビ番組に言及しているメッセージから，視聴者が番組中に注目している人物等を検出し，番組放送中に注目されている人や事柄をグラフで表す研究を行った．

まず，上原らは以下のアルゴリズムにより，番組で注目されている人や事柄をグラフで表している．また，上原らの作成したシステムの概要を図 16 に示す．

step1. 掲示板サイトから，番組放送中に書き込まれたメッセージのデータを取得する．

step2. メッセージを形態素解析し，代名詞や助詞等の汎用語を除外する．また，番組の出演者のニックネーム等の同意語を集約し，同意語辞書を作成する．

step3. 一定の時間間隔以上の頻度で出現する単語を注目語とする．注目語と判断した単語を，横軸が番組の放送時間で縦軸が出現頻度で表されるグラフに書き込む．

次に，上原らが行った実験について説明する．実験は，ドラマ番組において注目されている人物が誰であるのかをグラフで表し，表された人物が実際に画面に現れているかどうかで評価を行った．実験の結果，上原らの手法で注目されていると判断した人物は実際に画面上に登場していることを確認している．例外的に，注目されていると判断した人物が画面上に登場しておらず，他の出演者のセリフで話題になっているケースを確認している．

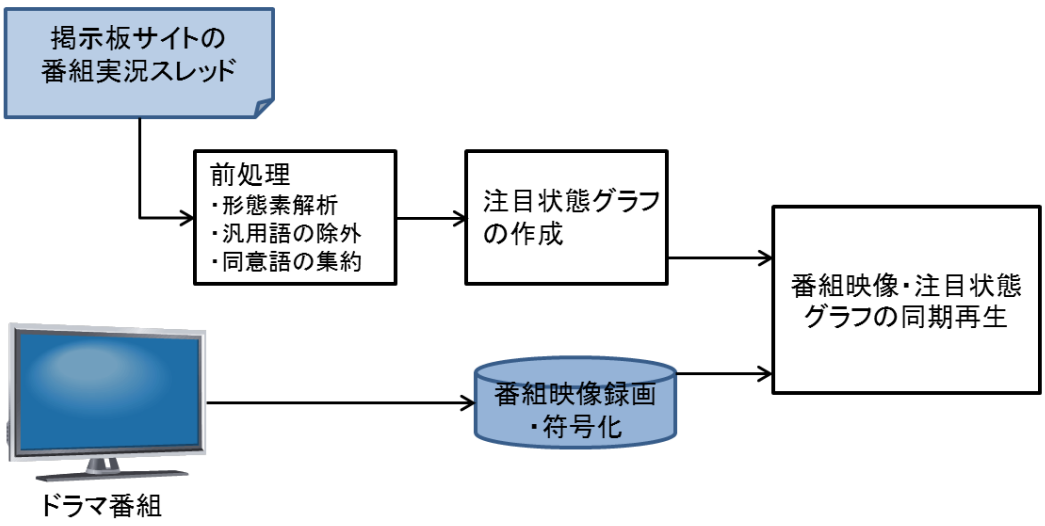

図 16　システムの流れ([24]の図 4 を参考に作成)

3. 2. 3　テレビ番組についてリアルタイムに書き込まれるコメントを対象とした研究のまとめ

テレビ番組についてリアルタイムに書き込まれるコメントを対象とした研究について述べた．本節で紹介した研究を表 3 にまとめる．

表 3 テレビ番組についてリアルタイムに書き込まれるコメントを対象とした研究のまとめ

研究	研究内容	番組名の判定方法
宮森らの研究[23]	掲示板サイトを利用することにより，番組の盛り上り場面や特定の視聴者が興味を示している場面を抽出し，ダイジェスト視聴に役立てる研究	2 ちゃんねる[1]に設置されている番組実況専用の掲示板を使用
上原らの研究[24]	掲示板サイトを利用することにより，番組で注目されている人物や事柄をグラフで表す研究	2 ちゃんねるに設置されている番組実況専用の掲示板を使用

[1] 2 ちゃんねる
http://www.2ch.net/

3. 3　Twitter からテレビ番組に言及している投稿を検出する研究

本節では，本研究と最も関連する，Twitter からテレビ番組に言及している投稿を検出する研究について述べる．3. 3. 1 項で小林らが行った単語の出現頻度によりテレビ番組に言及している投稿を検出する研究[20]について述べ，3. 3. 2 項でソーシャルビューイングを行えるアプリケーションについて説明する．

3. 3. 1　小林らの研究[20]

2010 年に名古屋大学の小林らは，単語の出現頻度によりテレビ番組を見ているユーザを発見し，そのユーザのテレビ番組に言及した投稿を検出する研究を行った．まず，図 17 に小林らの手法のシステムの流れを示す．小林らの手法は 2 段階からなる．1 段階目でテレビ番組を観覧しているユーザかどうかのラベル付けをし(図 17 の①)，2 段階目でテレビ番組に言及している投稿かどうかを判断する(図 17 の②).

1 段階目の詳細について説明する．1 段階目は次の 4 つステップにより行われる．

step1. 番組開始直後の投稿を抽出する．

step2. テレビ番組の開始直後は，「～がはじまった.」等の，番組開始を象徴する投稿が多く見られる．そこで，step1 で抽出した投稿を形態素解析し，番組開始直後の投稿に含まれる単語の出現頻度を学習する．

step3. ユーザが何らかのテレビ番組を見ているか否かを判断する SVM 識別器を作る．

step4. 番組開始直後にメッセージを投稿したユーザに「観覧」または「不観覧」のラベルを付加する．

次に，2 段階目の詳細について説明する．2 段階目は次の 4 つステップにより行われる．

step1. 1 段階目で「観覧」または「不観覧」のラベルを付加したユーザが投稿した，番組放送時間帯におけるその後のメッセージを抽出する．

step2. step1 で抽出した投稿を形態素解析し，特定の品詞の単語のみを残す．

step3. step2.で残した単語の出現頻度を学習して SVM 検出器を作成する．

step4. 番組放送時間帯に投稿されたメッセージに対し, step3.で作成した検出器を用いて，テレビ番組に言及している投稿を検出する．

小林らが行った評価実験について説明する．実験では，小林らの手法の 1 段階目を評価する実験と，2 段階目を評価する実験の 2 つを行っている．1 段階目を評価する実験では，テレビ番組 5 タイトルについて，放送開始後の 5 分間に Twitter に投稿された 614 個のメッセージを投稿したユーザに対し，「観覧」または「不観覧」のラベル付けする精度を測る実験を行った．まず，各投稿を形態素解析し，出現頻度の高い単語 200 語を要素とする 200 次元の特徴ベクトルを作成する．そして，人手により各投稿を行ったユーザに対して，「観覧」または「不観覧」のラベルを付与した正解セットを作成する．作成した特徴ベクトルを SVM により識別し，メッセージを投稿したユーザへの「観覧」または「不観覧」のラベル付けを行う．この時，SVM によるラベル付けの学習および評価は，3 hold cross-validation で行った．実験の結果，ユーザへのラベル付けの精度は平均して 90.1%であった．

次に，2 段階目を評価するために行われた実験について説明する．テレビ番組 1 タイトルについて，1 段階目でラベルを付加したユーザが，番組放送時間帯に投稿した 814 個のメッセージを用いて，実際に番組に言及している投稿を絞り込み，番組に言及している投稿の検出精度を測る実験を行った．1 段階目でラベルが付けられたユーザの番組放送中に投稿されるメッセージを形態素解析し，以下の 3 つの手法により，出現頻度の高い単語 500 語を要素とする 500 次元の特徴ベクトルを作成する．

手法1.	名詞のみ
手法2.	名詞，動詞，助動詞
手法3.	名詞，動詞，助動詞，形容詞，形容動詞

そして，番組に言及している投稿か否かの正解セットを人手で作成する．作成した特長ベクトルを SVM により識別し，番組放送時間帯に Twitter に投稿されたメッセージから，実際に番組に言及した投稿を検出する．この時，SVM による番組に言及した投稿検出の学習および評価は，3 hold cross-validation で行った．実験の結果，手法 3 の検出精度が最も高

く，79.9%であった．

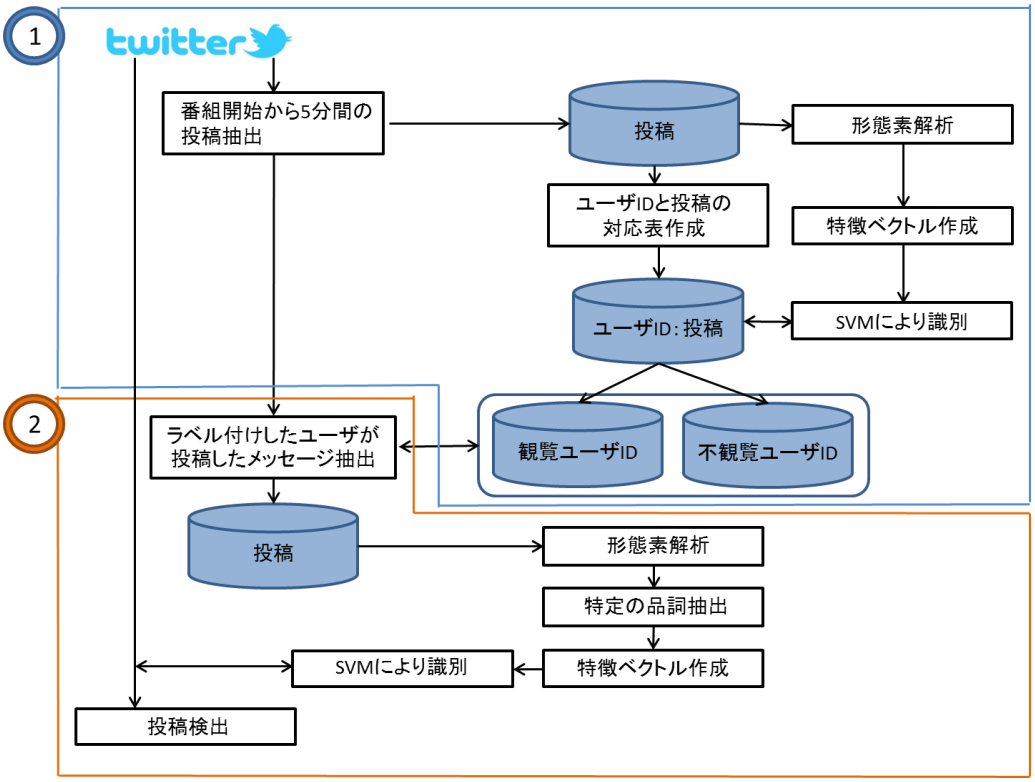

図 17　小林らが提案したシステムの流れ

3. 3. 2　ソーシャルビューイングを行えるアプリケーション

ソーシャルビューイングが普及する伴い，ソーシャルビューイングを容易に行えるアプリケーションが普及している．例えば，emocon[25]では，Twitter や Facebook 等の SNS 上の友人が見ているテレビ番組の感想を閲覧でき，自らも投稿を行えるスマートフォン用アプリケーションである．また，番組ごとに視聴している友人数及び番組の盛り上がりをグラフ化して表示する機能もあり，視聴する番組を選ぶ際の参考にすることも可能である．

また，テレ Viewing[26]では，emocon とは異なり SNS 上の友人に限定せず，放送中のテレビ番組に関する Twitter への投稿を閲覧できるスマートフォン用アプリケーションである．テレ Viewing においても，番組の盛り上がりがグラフで確認できる．また，テレ Viewing に対応しているテレビがあれば，アプリケーションから直接テレビのチャンネルを切り替える事もできる．emocon やテレ Viewing 等のソーシャルビューイングアプリケーションにおいて，表示される Twitter への投稿は，テレビ局のハッシュタグ及び番組のタイトルが含まれている投稿に限定されている．

図 18　emocon の画面(emocon 紹介ページ[2]より引用)

図 19 テレ Viewing の画面(テレ Viewing 紹介ページより引用)[3]

---

[2] グリーと VOYAGE GROUP が、テレビをより楽しむソーシャルビューイングアプリ“emocon”をリリース
http://app.famitsu.com/20121107_106749/

[3] ヤフー、テレビ番組のツイート状況が確認できるアプリ「テレ Viewing」
view-source:http://internet.watch.impress.co.jp/docs/news/20120724_548731.html

3. 3. 3　Twitter からテレビ番組に言及している投稿を検出する研究のまとめ

本節では, Twitter からテレビ番組に言及している投稿を検出する研究について紹介した. 本節で紹介した研究の手法と問題点を表 4 にまとめる.

小林らの手法の問題点について述べる．小林らの手法では，テレビ番組に関わらず汎用的なイベントについての投稿検出に適用可能である．しかし，小林らの研究では，番組終了後に検出器を作成するため，リアルタイムにテレビ番組についての意見を検出することができず，テレビ番組と Twitter のリアルタイム性を活用することが難しい.

次に，ソーシャルビューイングアプリケーションの問題点について説明する．ソーシャルビューイングアプリケーションでは，まず，テレビ局のハッシュタグが付けられている投稿を表示しているが，テレビ局のハッシュタグを使用するユーザは一部である．また，ハッシュタグはその機能上，タグが付けられた投稿を全て検出するため，感動を表現する投稿と番組中の人・物・事柄についての意見が混在してしまう．また，番組と関係のない投稿にハッシュタグを使用するユーザもいるため，意見投稿を検出することを考慮すると，ノイズとなる.

表 4　Twitter からテレビ番組に言及している投稿を検出する研究の手法と問題点

研究	手法	問題点
小林らの研究[20]	番組開始冒頭に投稿されたメッセージ中の単語の出現頻度を学習し，番組を見ているユーザ候補を作成する．そして，ユーザ候補の投稿中の単語の出現頻度を学習し，番組に言及している投稿を検出する	リアルタイムにテレビ番組についての意見を検出することができない
ソーシャルビューイングアプリケーション[25][26]	テレビ局のハッシュタグ及び番組のタイトル	・ハッシュタグを使用するのは一部のユーザ ・感情投稿や番組と無関係な投稿がノイズとなる

第4章　提案手法

本章では提案手法について述べる．本稿では，番組中の人・物・事柄等の名詞に言及したツイートを行っているユーザを意見投稿ユーザと定義し，意見投稿ユーザをリアルタイムに検出する手法を提案する．3. 3 節で述べたように，既存研究では，番組終了後に検出器を作成するため，リアルタイムにテレビ番組についての意見を検出することができず，テレビ番組と Twitter のリアルタイム性を活用することが難しい．また，ソーシャルビューイングアプリケーション等のハッシュタグによる検出では，意見投稿ユーザを見つけるという目的においては, 番組中の人・物・事柄等の名詞に言及していない『おおおおおおお #nhk』といった投稿や，番組と関係のない投稿がノイズとなる．

そこで，本稿では，意見投稿ユーザをのみをリアルタイム性を考慮し，検出することを目指す．リアルタイム性を考慮して検出することにより，ソーシャルビューイングを行なっているユーザが他人の意見を知る上で重要な投稿を閲覧でき，番組理解につがなる．また，テレビ局側にとっても，番組放送中に視聴者との双方向のコミュニケーションを行うことが可能となる．

提案手法では，まず，対象番組を選択した後，選択番組についての特徴語となる名詞を抽出しながら，番組放送時間帯に得られる特徴語を含んだ投稿を行うユーザを検出する．テレビ番組についての特徴語は，以下に示す 2 つの特徴語群を使用する．(1)電子番組表及びテレビ番組の字幕放送において表示される字幕テキストから得られる，番組公式の特徴語群，(2)Twitter への投稿からトピックモデルを利用して抽出される，SNS ユーザが生成する番組の特徴語群.

提案手法の詳細について 4. 1 節で番組公式の特徴語の抽出について述べ，4. 2 節で SNS ユーザが生成する番組の特徴語の抽出について説明する．そして，4. 3 節で特徴語群を利用した，意見を持ったユーザ検出について述べる．最後に 4. 3. 3 節で，番組公式の特徴語の抽出の際に必要となる，名詞の重要度の閾値の決定方法について説明する．

4. 1　番組公式の特徴語の抽出

テレビ番組公式の特徴語を取得する手順について述べる．まず，特徴語を取得する流れを図 20 に示す.

電子番組表については，予め形態素解析を行い，名詞を抽出しておく．次に，字幕テキストについては，字幕テキストが画面に表示された際，つまりテキストデータを取得した時点で形態素解析を行い，名詞を抽出する．そして，電子番組表及び番組開始から現時点

までに表示された字幕テキストから得られた名詞の重要度を計算し，特徴語を得る．

テレビ番組内で放送される話題は時系列的に変化し，それに伴いテレビ番組の特徴語も変化する．よって，新たな字幕テキストデータを取得した時点で，逐次的に名詞の重要度を再計算し，特徴語の更新を行う．

次に，名詞の重要度の計算について説明する．名詞の重要度を計算するにあたり，以下の 3 点を考慮し，番組の特徴語となる名詞を抽出する．

・現在までに放送された番組の中で特定の番組にのみ出現する名詞(idf 値が大きい名詞)
・字幕テキストに出現した名詞群の内、出現時刻が新しい名詞
・番組中に何度も字幕テキストに出現する名詞(tf 値が大きい名詞)

具体的には，名詞term_iの重要度$Importance_i$を以下の式(7)で定義し，最終的には，$Importance_i$が閾値以上となる名詞を番組の特徴語群として採用する．

$$Importance_i = tf_i \cdot idf_i \cdot e^{-\lambda t_i} \tag{7}$$

tf_iは名詞term_iが番組開始から現時点までに字幕テキストに出現した頻度であり式(8)で表される．

$$tf_i = \frac{n_i}{\sum_k n_k} \tag{8}$$

ただし，n_iは，対象番組中の字幕テキストにおける当該番組開始から現時点までの名詞term_iの出現回数であり，$\sum_k n_k$は，対象番組開始から現時点までに字幕テキストに出現した名詞の総数である．

idf_iは名詞term_iの逆文書頻度であり式(9)で表される．

$$idf_i = \log\frac{P}{p_i} \tag{9}$$

ただし，Pは過去に放送されたテレビ番組から現時点までの，字幕テキストが存在するテレビ番組の総数であり，p_iは字幕テキストに名詞term_iが出現するテレビ番組数である．

また，$e^{-\lambda t_i}$中のt_iは，名詞term_iが，字幕テキストに最後に出現してから現時点までの時間である．

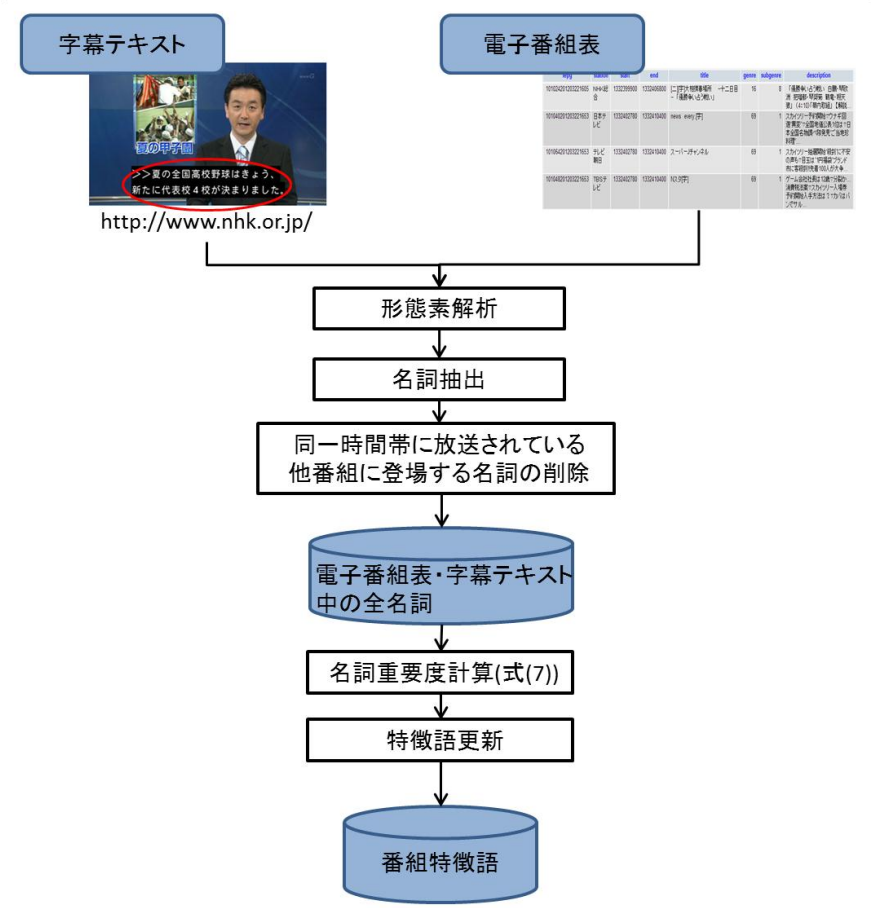

図 20　テレビ番組公式の特徴語を取得する流れ

4. 2　SNS ユーザが生成する番組の特徴語の抽出

本節では，SNS ユーザが生成する番組の特徴語を取得する手順について説明する．提案手法では，まず，Labeled LDA[21]における文書，単語，トピックラベルを，それぞれ，ツイート，単語，ハッシュタグに置き換えることにより，Labeled LDA を適用する．そして，Labeled LDA により求まる各トピックの単語分布を推定することにより，SNS ユーザが生成する番組の特徴語を抽出する．SNS ユーザが生成する特徴語として，以下に示す 3 種類の特徴語を想定している．これらの特徴語は，必ずしもテレビ番組中の人・物・事柄の正式な名称ではないが，テレビ番組中の人・物・事柄に着目しており，着目した対象に対しての意見を述べている投稿である可能性が高い．

1. 字幕中には登場しないが，ソーシャルビューイングを行うユーザが着目している固有名詞　(例：ドラマのキャストやアニメの声優)
2. 複数の意味を持つ単語が固有名詞として出現した場合の名詞　(例 : 大津(滋賀県大津市では一般名詞だが，サッカー選手の名称としては番組の特徴語となり得る．))
3. 字幕テキストには出現しない，テレビ番組中の人・物・事柄の別称.

特徴語を取得する流れを図 21 に示す．まず，Twitter への投稿からハッシュタグが付けられているすべての投稿を抽出する．次に，投稿からハッシュタグを除き，除いたハッシュタグを投稿に対するラベルとする．そして，残りの投稿文を形態素解析し，名詞のみを残す．つまり，Twitter への投稿$d \in D$(Dは tweet の全体集合)に付加されているハッシュタグ及び名詞をそれぞれ，ハッシュタグの有無を表すラベル列$\boldsymbol{\Lambda}^{(d)} = \{l_1, \ldots, l_K\}$(Kはハッシュタグの集合)と名詞リスト$\boldsymbol{w}^{(d)} = \{w_1, \ldots, w_{N_d}\}$とする．ただし，$w_i \in V$(Vは名詞の語彙集合)，$l_k \in \{0,1\}$である．そして，Twitter への投稿集合Dに対して Labeled LDA を適用することにより推定されるハッシュタグkに対する名詞分布$\widehat{\boldsymbol{\beta}}_k$を得る．そして，検出対象とした番組に関連するハッシュタグkを選び，kから生成される確率の高い名詞を，特徴語として抽出する．

3.1 節で述べたように，テレビ番組内で放送される話題は時系列的に変化し，それに伴いテレビ番組の特徴語も変化するため，5 分毎に，直近 5 分間の投稿に対し，上記の手順により SNS ユーザが生成する番組の特徴語を取得する．

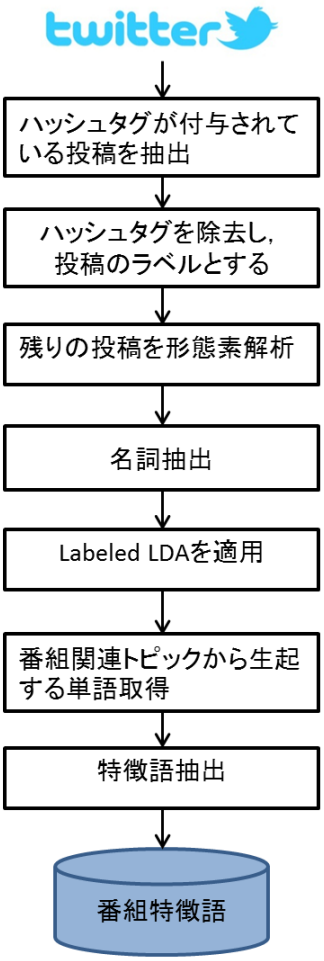

図 21　SNS ユーザが生成する特徴語を取得する流れ

4.3　特徴語群を利用した，意見を持ったユーザ検出

本節では，意見を持ったユーザを検出する方法について説明する．第 4 章の冒頭で述べたように，検出の対象とする意見を持ったユーザの定義は，番組中の人・物・事柄等の名詞に言及しているユーザとした．番組中の人・物・事柄等の名詞を含んだ投稿を行うユーザは，その投稿中で番組中の人・物・事柄等の感想や意見を述べている可能性が高い．よって，まず検出対象とした番組の放送時間において，Twitter に投稿されたメッセージ中に，番組の特徴語群が含まれている投稿を検出する．そして，検出された投稿を行ったユーザを，投稿検出のきっかけとなった番組の特徴語に対する意見を持ったユーザとして検出する．ここで，検出した投稿を行ったユーザは，2. 1 節の表 1 に示した”screen_name”により得ることができる．

まず，4. 3. 1 項で番組公式の特徴語群を用いた検出手法の説明を行い，次に，0 項で SNS ユーザが生成した特徴語群を用いた検出手法の説明を行う．最後に 4. 3. 3 項で両者を統合した,『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』の提案手法について述べる．

4. 3. 1　番組公式の特徴語群を用いた検出手法

番組公式の特徴語を用いた検出手法の方法について説明する．まず，検出の流れを図 22 に示す．

番組公式の特徴語群を用いた検出手法では，検出対象としたテレビ番組の放送時間帯において，Twitter API により投稿を取得した時点で，その投稿中に現時点で番組公式の特徴語群となっている語が含まれているかをチェックすることにより，テレビ番組への意見投稿として検出する．そして，検出された投稿を行ったユーザを，投稿検出のきっかけとなった番組の特徴語に対する意見を持ったユーザとして検出する．

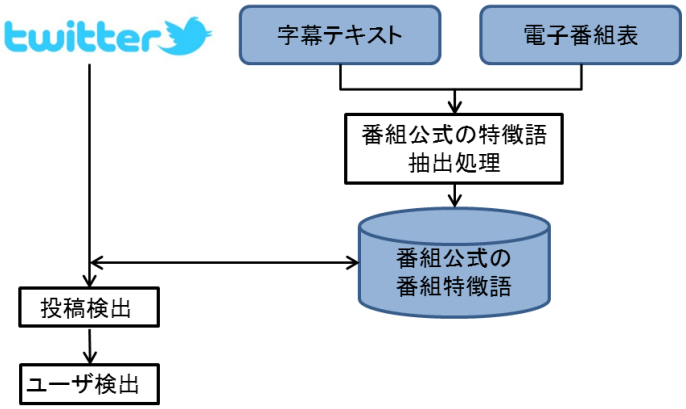

図 22　番組公式の特徴語群を用いた検出手法の流れ

4. 3. 2　SNS ユーザが生成した特徴語群を用いた検出手法

SNS ユーザが作成した特徴語群を用いた検出手法の方法について説明する．まず，検出の流れを図 23 に示す．

SNS ユーザが作成した特徴語群を用いた検出手法では，検出対象としたテレビ番組の放送時間帯において，5 分毎に SNS ユーザが作成した番組特徴語の抽出処理と番組へ意見投稿を行うユーザ検出を行う．Twitter API により取得した 5 分間の投稿を用いて，SNS ユーザが作成した特徴語抽出処理を行い，抽出処理によって得られた特徴語が，取得した 5 分間の投稿に含まれているかをチェックすることにより，テレビ番組への意見投稿として検出する．そして，検出された投稿を行ったユーザを，投稿検出のきっかけとなった番組の特徴語に対する意見を持ったユーザとして検出する．

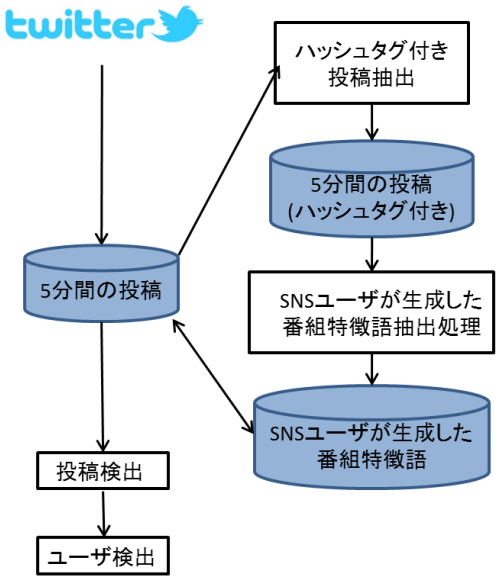

図 23　SNS ユーザが作成した特徴語を用いた検出手法の流れ

4. 3. 3　提案手法

本項では，『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』の提案手法について説明する．提案手法では，4. 3. 1 項で説明した番組公式の特徴語を用いたユーザ検出手法と，4. 3. 2 項で説明した SNS ユーザが作成した特徴語群を用いたユーザ検出手法を併用し，ユーザの検出を行う．提案手法の流れを図 24 に示す．

番組公式の特徴語群を用いたユーザ検出と SNS ユーザが生成した特徴語群を用いたユーザ検出では，検出結果が重複する場合があるため，結果を統合するときに重複を削除し，最終的な検出結果とする．

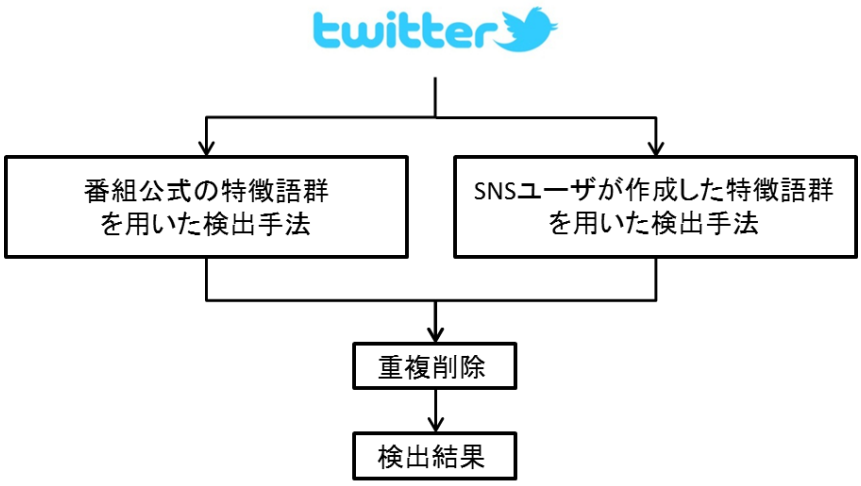

図 24　提案手法の流れ

4.4　重要度$Importance$の閾値決定方法

本節では，番組の特徴語とする重要度$Importance$の閾値を求める方法ついて説明する．一般的に閾値を低く設定し過ぎると一般的な名詞を除外することができず，テレビ番組についての意見でない投稿を検出する可能性が高まる．また，閾値を高く設定し過ぎると番組の特徴となるべき名詞を特徴語群にすることができない場合があり，番組についての意見を検出できなくなる．

重要度$Importance$の閾値の決定の方法について説明する．まず，重要度$Importance$の閾値を$Threshold_{imp}$と表す．また，正解 tweet 集合はテレビ局のハッシュタグ及びテレビ番組自体のハッシュタグが付けられた投稿から以下に示す投稿を人手により削除した tweet 集合とする．

(1) 名詞が含まれておらず，ハッシュタグなしには投稿の主語を推測出来ない，感動を表現する投稿
例：うおおおおおおおおおおおおおおおお #nhk

(2) テレビ番組とは無関係にハッシュタグを使用している投稿
例：政府の悪法オリンピック、開催中(￣□￣;)!! 種目：【 #ACTA #TPP #違法 DL 刑罰化 #秘密保全法案】 #オリンピック #Olympic2012 #nhk #tbs #tvasahi #fujitv #ntv

重要度$Importance$の閾値の決定は以下に示す方法により算出する．

step1. 閾値決定に用いるデータセットに対し，$Threshold_{imp}$を変化させながら，提案手法を用いて投稿を検出する．

step2. step1.による投稿検出の再現率$Recall$(式(10))及び適合率$Precision$(式(11))を算出する．

$$Recall = \frac{\text{step1. で検出できた正解 tweet 集合}}{\text{正解 tweet 集合}} \quad \textbf{(10)}$$

$$Precision = \frac{\text{step1. で検出できた正解 tweet 集合}}{\text{step1. で検出したハッシュタグが付いている全ての tweet 集合}} \quad \textbf{(11)}$$

step3. step2 で求めた$Recall$と$Precision$により$F-measure$(式(12))を計算する．

$$F-measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad \textbf{(12)}$$

step4. $F-measure$が最大となる閾値を，番組検出を行うジャンルにおけるテレビ番組の最終的な$Threshold_{imp}$とする．

第5章　実験・評価

本章では，提案したソーシャルビューイングを行なっている Twitter ユーザから，意見を持ったユーザを発見する手法の評価実験と結果について説明する．5. 1 節で使用データについて説明し，5. 2 節で評価方法を述べる．そして，5. 3 節で重要度*Importance*の閾値決定方法について説明し，5. 4 節で評価結果についてまとめる．

5. 1　使用データ

実験に用いる電子番組表及び字幕テキストは，2005 年 1 月から 2010 年 11 月及び，2012 年 3 月 22 日から 2013 年 1 月 12 日までの，関東で放送された 7 局のテレビ番組における字幕テキストである．字幕テキストの収集期間中において，字幕テキストが存在していたテレビ番組数は 258,342 番組である．

Twitter のデータは, 2011 年 12 月 3 日から 2013 年 1 月 3 日まで Twitter API により収集したデータおよび 2012 年 8 月 2 日から 2012 年 8 月 14 日までに Gnip 社[27]の API により収集したデータを使用した．実験に使用したデータを表 5 にまとめる．

評価実験では，電子番組表及び字幕テキストを収集できた番組の中から，意見を投稿するユーザを検出する対象のテレビ番組を複数選択し，選択したテレビ番組の電子番組表及び字幕テキストと, 番組放送時間帯に Twitter に投稿されたメッセージのデータを利用して, 5. 2 節で説明する実験を行った．

表 5 実験に使用したデータ

	収集期間	データ量
電子番組表及び字幕テキスト	2005 年 1 月から 2010 年 11 月及び 2012 年 3 月 22 日から 2013 年 1 月 12 日	258,342 番組
Twitter のデータ	Twitter API：　2011 年 12 月 3 日から 2013 年 1 月 3 日	380 日分
	Gnip 社 API：2012 年 8 月 2 日から 2012 年 8 月 14 日	13 日分

5.2 評価方法

本節では，提案手法の評価実験について説明する．実験対象として，アニメ，ドキュメンタリー，映画，スポーツの計 4 ジャンルから各々1 番組を選んだ．そして，ジャンルごとに以下に示す手順により実験を行った．また，表 6 に実験で使用したデータを，表 7 に実験に用いた Labeled LDA のパラメータをまとめる．Labeled LDA のパラメータは[28]を参考に定めた．ここで，step4.における番組に関連するハッシュタグは，番組が放送されているテレビ局のハッシュタグ及び番組自体のハッシュタグとした．

また，step6.において算出する適合率は以下に示す式(13)を用いる．適合率を式(13)のように定める理由は，テレビ番組に意見投稿を行うユーザがハッシュタグを使用する際，必ずしもテレビ局のハッシュタグまたは番組自体のハッシュタグを使用するとは限らないからである．例えば，表 6 の番組 D(スポーツ)は，ロンドンオリンピック・男子サッカーについて放送された番組であり，正解セットとしたハッシュタグ以外にも，#soccer，#olympic，#オリンピック，#jfa(Japan Football Association を意味する)，#U23(23 歳以下日本代表を意味する)等のハッシュタグを使用して，番組について投稿されている．番組について意見している投稿は，(1) 番組の出演者やキャストを含む人・物・事柄や番組自体について言及している投稿，(2) 番組の内容に即している投稿の 2 つの基準を共に満たす投稿を人手により抽出した．

$$Precision = \frac{\text{分母に含まれる投稿のうち，番組について意見している投稿}}{\text{検出した投稿からランダムに選んだ 100 件の投稿}} \qquad (13)$$

step1. 実験を行う番組を選択する．

step2. step1.で選択した番組放送時間帯に Twitter に投稿されたメッセージを取得する．

step3. 取得したデータを 3 分割し，そのうちの 1 つを用いて，4. 3. 3 節で説明した方法により，重要度$Importance$の閾値を決定する．

step4. step3.で閾値決定に使わなかったデータセットをデータセット A，データセット B とし，それぞれのデータセット中から番組に関連するハッシュタグが付加されている投稿を抽出する．

step5. step4.でデータセット A から抽出した投稿において，以下に示す投稿を人手により削除し，正解セット A とする．同様の処理をデータセット B にも行い，正解セット B を作成する．

(1) 名詞が含まれておらず，ハッシュタグなしには投稿の主語を推測出来ない，感動を表現する投稿
例：うおおおおおおおおおおおおおおおお #nhk

(2) テレビ番組とは無関係にハッシュタグを使用している投稿
例：政府の悪法オリンピック、開催中(￣□￣;)!! 種目：【 #ACTA #TPP #違法 DL 刑罰化 #秘密保全法案】 #オリンピック #Olympic2012 #nhk #tbs #tvasahi #fujitv #ntv

step6. データセット A，B に対し，(1) 番組公式の特徴語を用いたユーザ検出，(2) SNS ユーザが作成した特徴語群を用いたユーザ検出を行い，適合率(式(13))，再現率(式(10))及び$F-measure$(式(12))を求め，それぞれの平均値を算出する．
ただし，(2) SNS ユーザが作成した特徴語群を用いたユーザ検出では，データセット A(B)から得られる SNS ユーザが作成した特徴語群を用いて，データセット B(A)に対しユーザの検出を行う．また，番組に関連するハッシュタグが付けられた投稿から生起する確率の高い Top 5 の単語を SNS ユーザが作成した特徴語とする．

step7. step6.でデータセット A(B)に対し，2 つの手法を適用し検出した投稿の和集合を取り，提案手法の適合率(式(13))，再現率(式(10))及び$F-measure$(式(12))を求め，それぞれの平均値をとる．

表 6 実験に使用した番組と Twitter のデータ

	番組 A (アニメ)	番組 B (ドキュメンタリー)	番組 C (映画)	番組 D (スポーツ)
番組の放送日時	2011/12/9 21:00 ～ 2011/12/9 23:30	2012/8/6 20:00 ～ 2012/8/6 20:50	2012/8/10 21:00 ～ 2012/8/10 22:54	2012/8/8 0:25 ～ 2012/8/8 3:00
番組放送時間帯に API で取得した Twitter に投稿されたメッセージ数	81,781	371,730	988,061	1,035,134
正解セットとしたハッシュタグ	#ntv, #laputa	#nhk, #nhkspecial	#ntv, #チャーリーとチョコレート工場	#tbs, #daihyo
正解セットとしたハッシュタグが付けられていた投稿数 (step4.で抽出される投稿数)	3,783	228	3,335	5,533
正解セットとした投稿数 (step5.で抽出される投稿数)	1,258	171	2,578	3,748

表 7 実験に使用した Labeled LDA のパラメータ

α	η	Gibbs Sampling におけるイテレーション数
0.01	0.01	1,500

5. 3　重要度$Importance$の閾値決定

本節では，各ジャンルの番組における重要度$Importance$の閾値決定のための事前実験について説明する．

4. 3. 3 節で説明したように，$Threshold_{imp}$を 0.000 から 0.020 まで 0.001 刻みで変化させ，各ジャンルの番組において$F-measure$が最大となる$Threshold_{imp}$を求める．アニメにおける$F-measure$の値，ドキュメンタリーにおける$F-measure$の値，映画における$F-measure$の値，スポーツにおける$F-measure$の値をそれぞれ図 25～図 28 に示す．実験の結果より，各ジャンルにおける番組において，最終的に採用する$Threshold_{imp}$の値を表 8 にまとめる．

表 8 各ジャンルの番組の$Threshold_{imp}$

	番組 A (アニメ)	番組 B (ドキュメンタリー)	番組 C (映画)	番組 D (スポーツ)
$Threshold_{imp}$	0.0090	0.016	0.017	0.0050

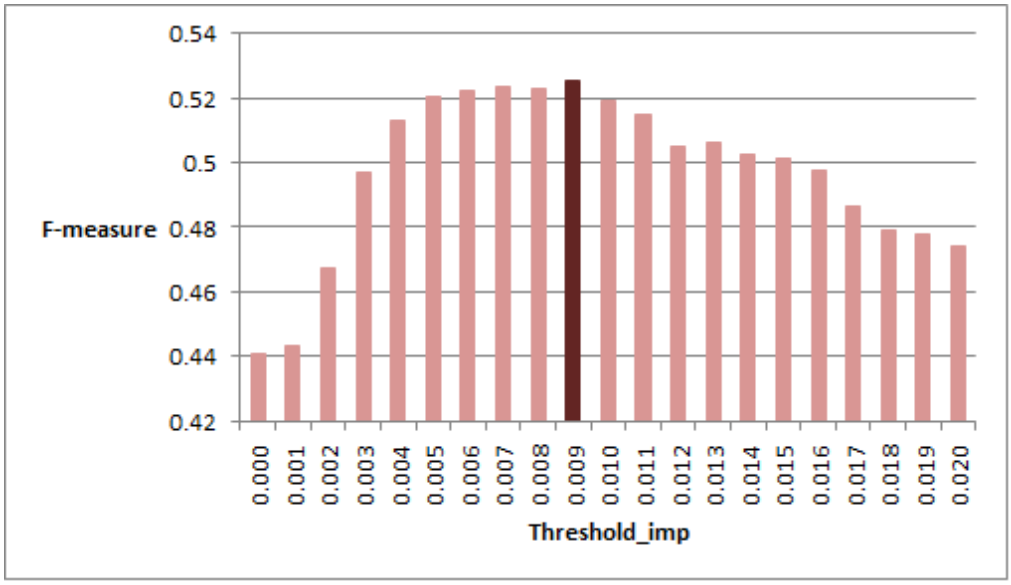

図 25 アニメにおける$F-measure$の値

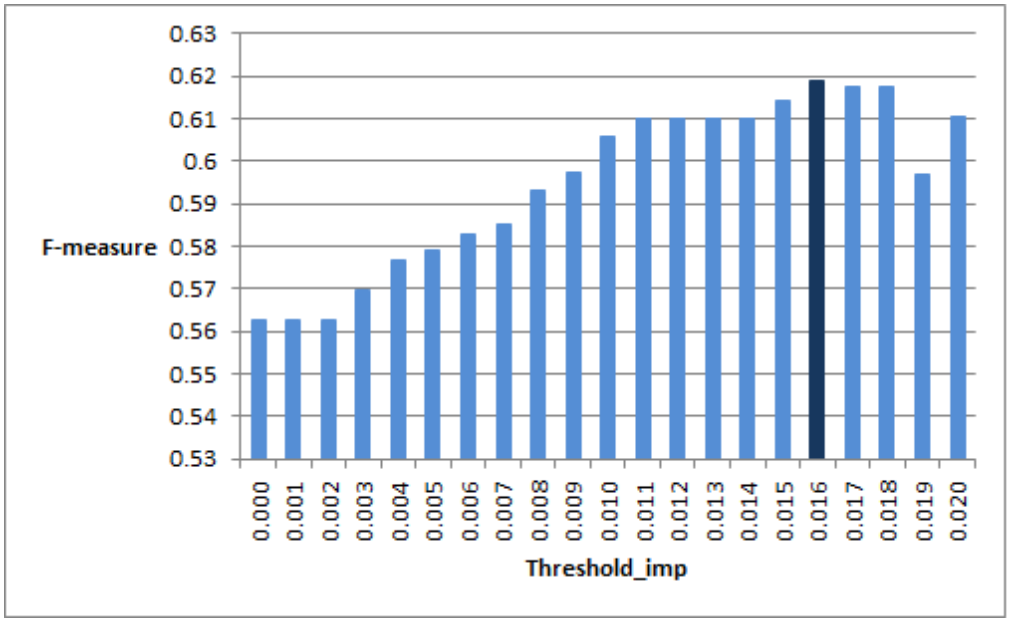

図 26　ドキュメンタリーにおける$F-measure$の値

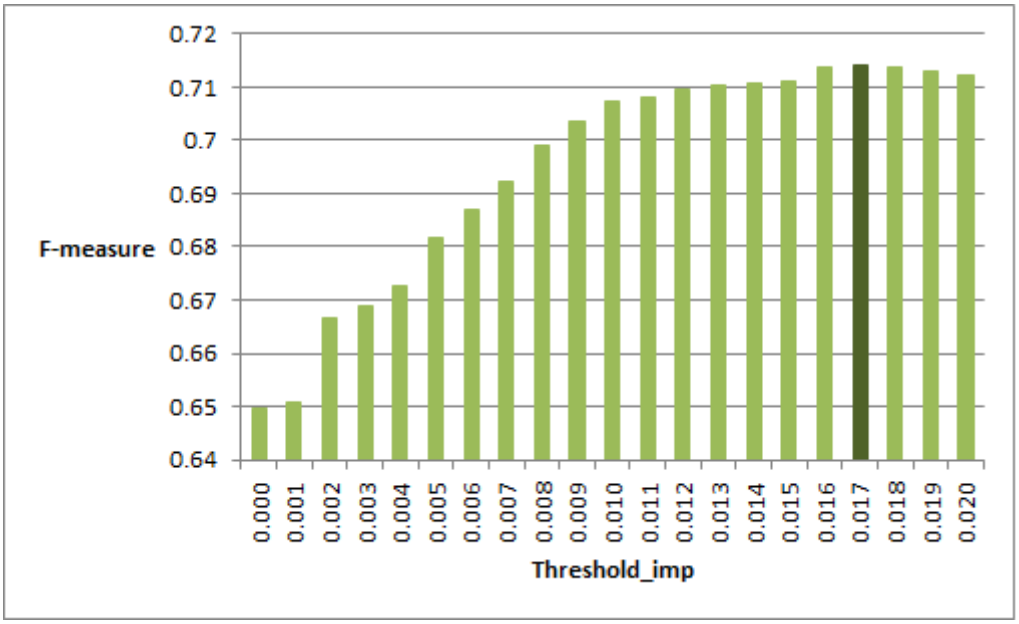

図 27　映画における$F-measure$の値

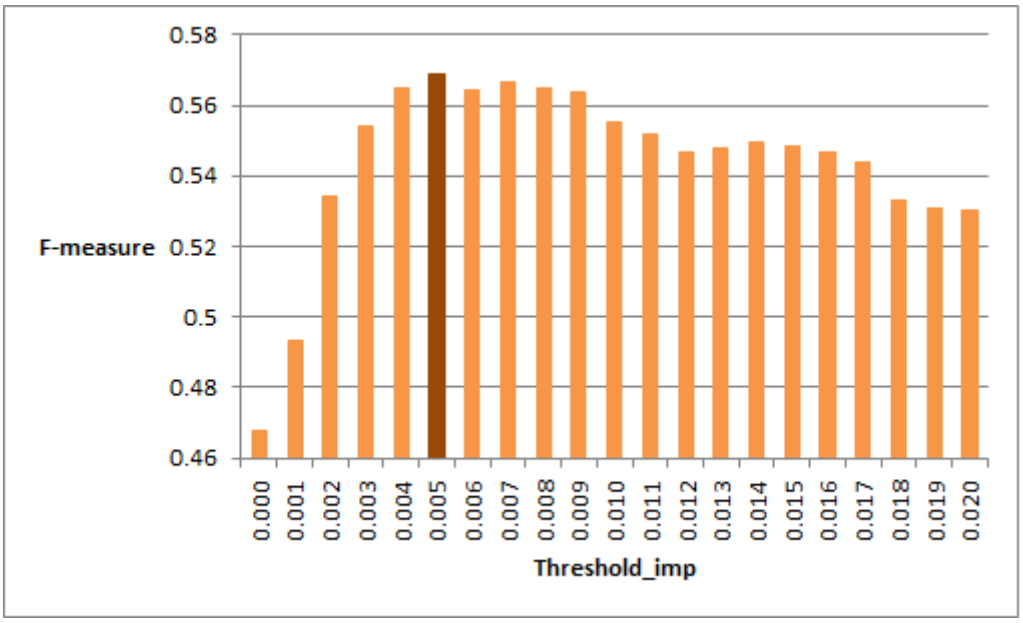

図 28　スポーツにおける$F-measure$の値

5. 4　評価結果

本節では，評価実験の結果について説明する．まず，5. 4. 1 項で実験結果の総評を行う．次に，5. 4. 2 項で実験の結果分かったハッシュタグの使用率について述べる．そして，5. 4. 3 で誤検出について考察し，5. 4. 4 項でテレビ番組についての意見投稿を行ったユーザを検出できなかった原因について考察する．最後に，番組公式の特徴語群と SNS ユーザが生成した特徴語群の違いについて考察を述べる．

5. 4. 1　実験結果の総評

本項では，評価実験の結果について述べる．まず，意見投稿検出の実験の結果得られた適合率を図 29 に，再現率を図 30 に，$F-measure$を図 31 に示す．実験の結果，提案手法は平均して 76%の適合率を保ちながら 68%の再現率で意見投稿を検出することができた．

4 番組に共通して見られる傾向としては，以下に示す 2 つの傾向が挙げられる．

1. 適合率と再現率共に，番組公式の特徴語による検出が，SNS ユーザが生成する特徴語のみによる検出を上回っている．
2. 番組公式の特徴語による検出と，SNS ユーザが生成する特徴語のみによる検出を併用することにより，再現率が向上している．

傾向 1 について考察すると，適合率については，番組公式の特徴語群は，idf 値により一般名詞が特徴語として入ることが少ためと考えられる．また，再現率については，SNS ユーザが生成する特徴語数に制限を設けた一方，字幕テキスト及び電子番組表を用いているため，番組の特徴語の種類が多ためと考えられる．

傾向 2 については，番組公式の特徴語による検出が，SNS ユーザが生成する特徴語のみによる検出を併用することにより，より多くの特徴語が得られ，片方の手法のみでは検出できない投稿を検出できていることがわかる．この傾向については，5. 4. 5 項で考察する．

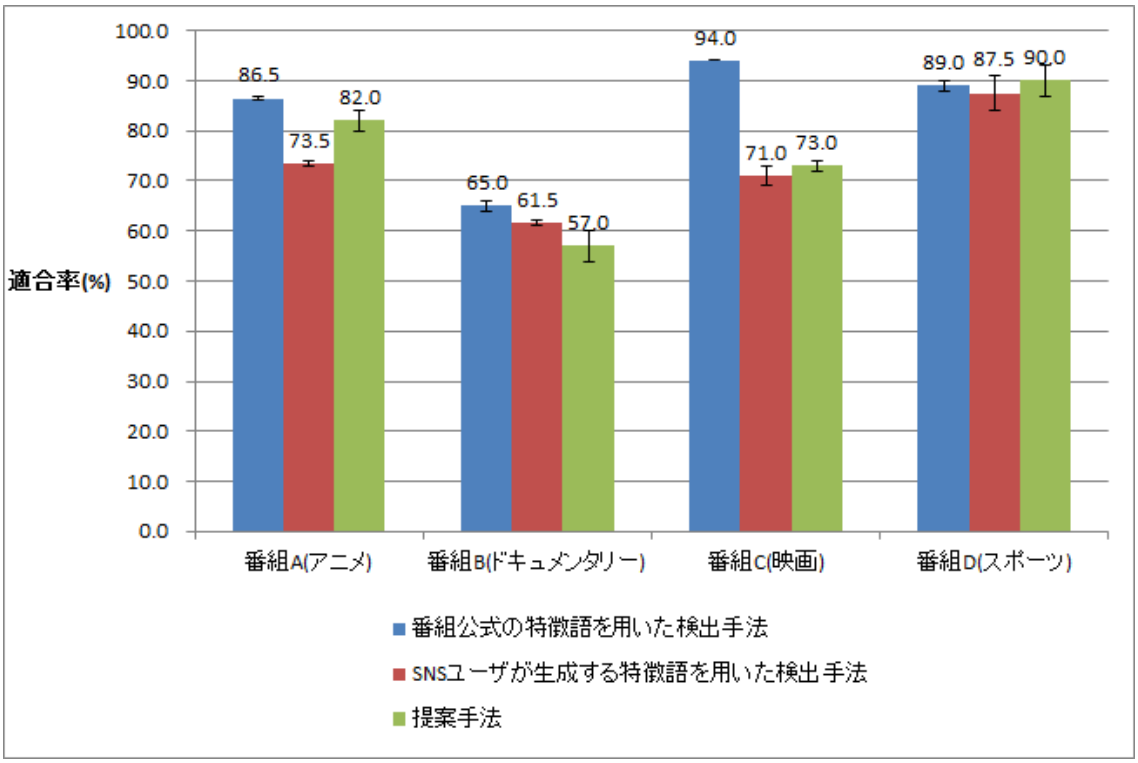

図 29　意見投稿検出の実験結果における適合率

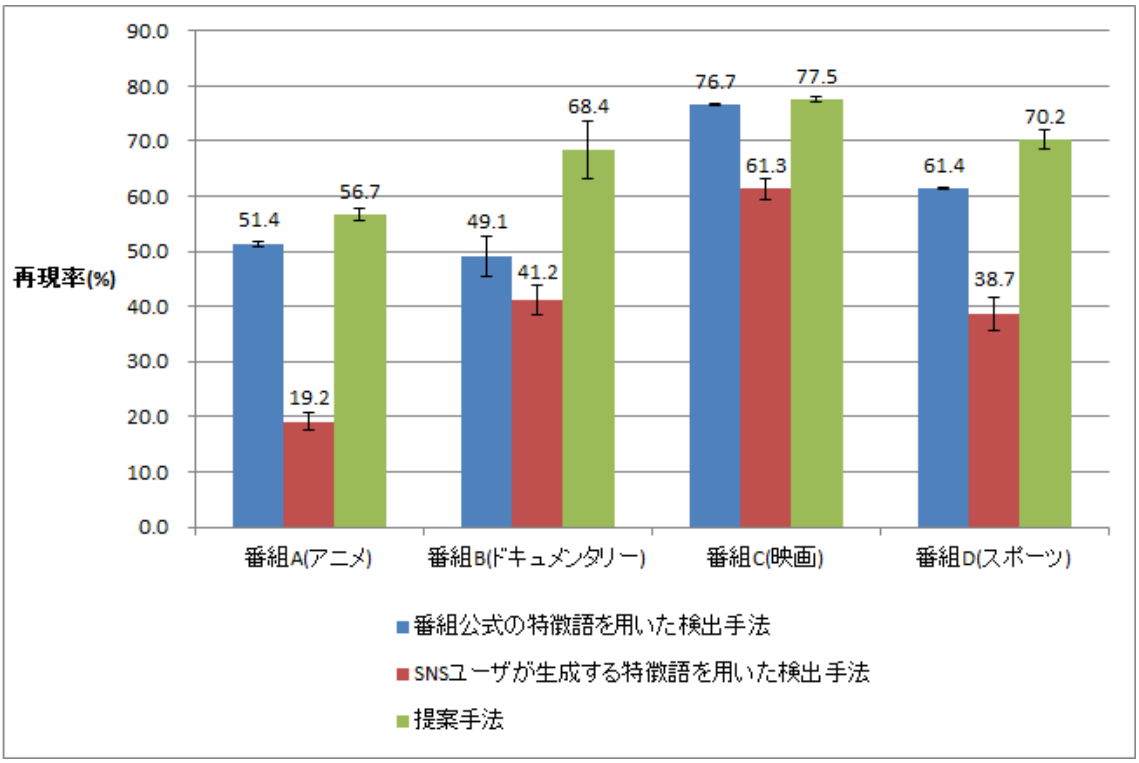

図 30　意見投稿検出の実験結果における再現率

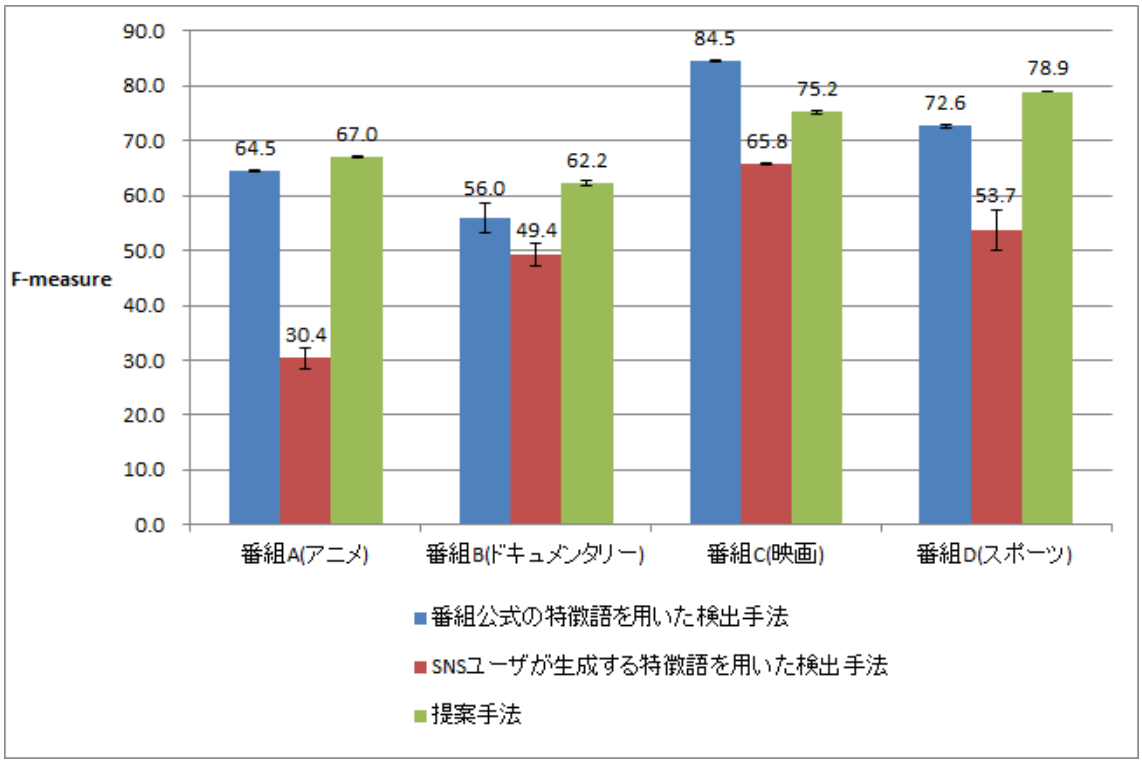

図 31　意見投稿検出の実験結果における *F-measure*

5. 4. 2　ハッシュタグの使用率

提案手法では，テレビ局やテレビ番組自体のハッシュタグが付けられていないユーザであっても，検出可能である．そこで，検出できた番組についての意見投稿中に含まれるテレビ局やテレビ番組自体のハッシュタグの使用率を調査した．

この調査では，5. 2 節において，提案手法による投稿検出の適合率(式(13))を算出する際に用いた，ランダムサンプリングされた投稿を使用した．提案手法による投稿検出はデータセット A，B の 2 つで行ったため，計 200 件の投稿を使用する．

まず，ランダムサンプリングされた 200 件の投稿から，適合率(式(13))を算出する際と同様に，番組について言及している投稿は，(1) 番組の出演者やキャストを含む人・物・事柄についての投稿，(2) 番組の内容に即している投稿の 2 つの基準を共に満たす投稿を人手により抽出した．つまり，200 件の投稿から，実際に番組について意見を言及している投稿を抽出した．そして，抽出した投稿中に，表 6 の正解としたハッシュタグがどれだけの割合で含まれているかを確認した．その結果を図 32 に示す．

図 32 より，テレビ番組について意見を行うユーザのうち，テレビ局やテレビ番組自体のハッシュタグを使用するユーザは一部であることがわかる．また，提案手法は，テレビ局やテレビ番組自体のハッシュタグを使用していないユーザの意見投稿を検出できている

ことが確認できた.

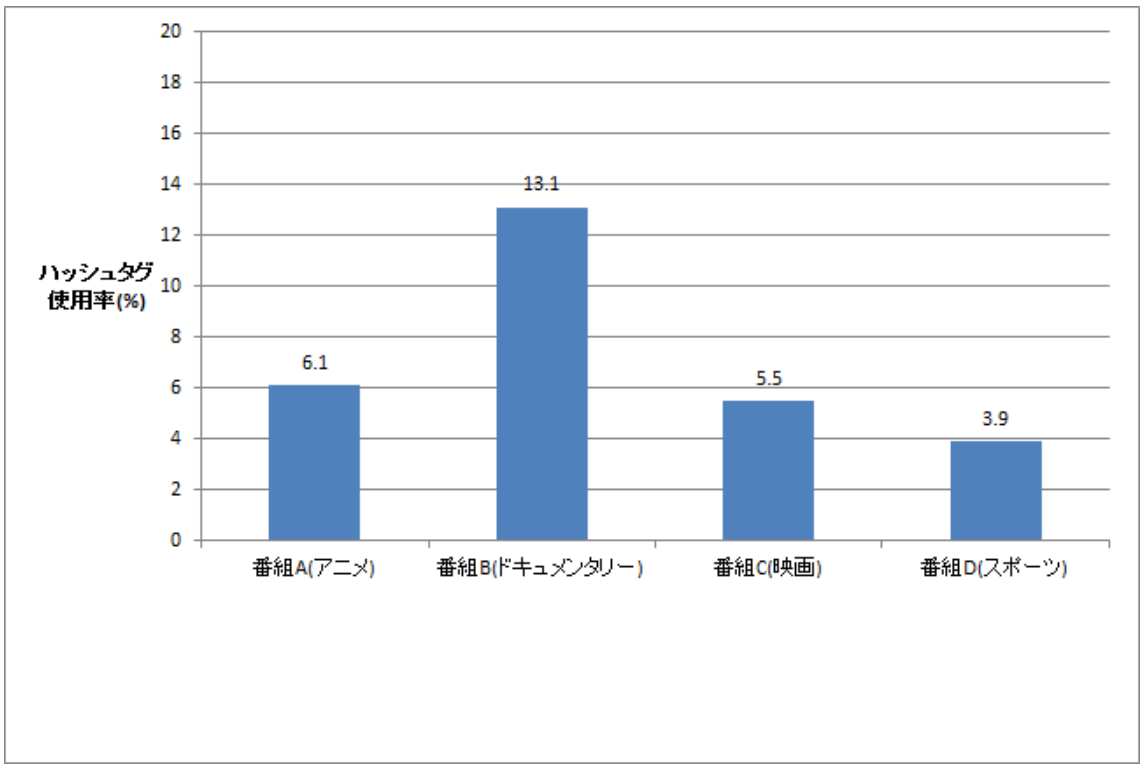

図 32　意見投稿中のテレビ局のハッシュタグ及び番組のハッシュタグ使用率

5. 4. 3 誤検出についての考察

本項では，誤検出について考察する．まず，誤検出となる投稿は以下に示すように 3 つに分類することができた．ここで，3. 番組に登場する人・物・事柄をパロディにしている投稿とは，番組 A(アニメ)に対する『あなたを天空の城ラピュタの登場キャラで表すと… http://t.co/UEIcktEU です。 http://t.co/xV38JpoT #lapu_chara だれ!?(笑)』のような投稿のことである．

1. 番組の特徴語を含んでいるが，番組には言及していない投稿
2. 番組を録画しているまたは番組を見られていない事を示す投稿
3. 番組に登場する人・物・事柄をパロディにしている投稿

誤検出のとなる原因の割合をテレビ番組のジャンルごとにまとめたものを図 33 に，検出手法ごとにまとめたものを図 34 に示す．

番組ジャンル別による誤検出の原因では，4 番組に共通して「1.番組の特徴語を含んでいるが，番組には言及していない投稿の割合」が一番大きくなった．これは，番組の特徴語とした語が，一般的な名詞と成り得る可能性が大きいことを表している．

また，番組別の特徴としては，番組 B(ドキュメンタリー)には，「2. 番組を録画しているまたは番組を見られていない事を示す投稿」，「3. 番組に登場する人・物・事柄をパロディにしている投稿」が見られなかった．これは，番組中で扱っている内容が原子爆弾についてであり，番組の内容をパロディにすることはできなかったためと考えられる．一方，番組 A(アニメ)は，「2. 番組を録画しているまたは番組を見られていない事を示す投稿」の割合が比較的高くなった．番組 A(アニメ)は非常に人気の番組であり，誤検出した投稿中に番組 A(アニメ)を見られないことを悔やむ投稿が多く見られた．さらに，番組 C(映画)は，ファンタジー映画であり，パロディを作りやすかったことから，「3. 番組に登場する人・物・事柄をパロディにしている投稿」の割合が比較的大きくなったと考えられる．

次に，検出手法ごとの違いについて考察する．検出手法別の誤検出の原因においても，4 番組に共通して「1.番組の特徴語を含んでいるが，番組には言及していない投稿」の割合が一番大きくなった．SNS ユーザが生成する特徴語のみによる検出は，番組公式の特徴語のみによる検出と比べ，「1.番組の特徴語を含んでいるが，番組には言及していない投稿」の割合が高かった．図 29 に示した投稿検出の適合率と共に考察すると，「1.番組の特徴語を含んでいるが，番組には言及していない投稿」による誤検出の絶対数も大きく，適合率の低下の要因になったのではないか考えられる．

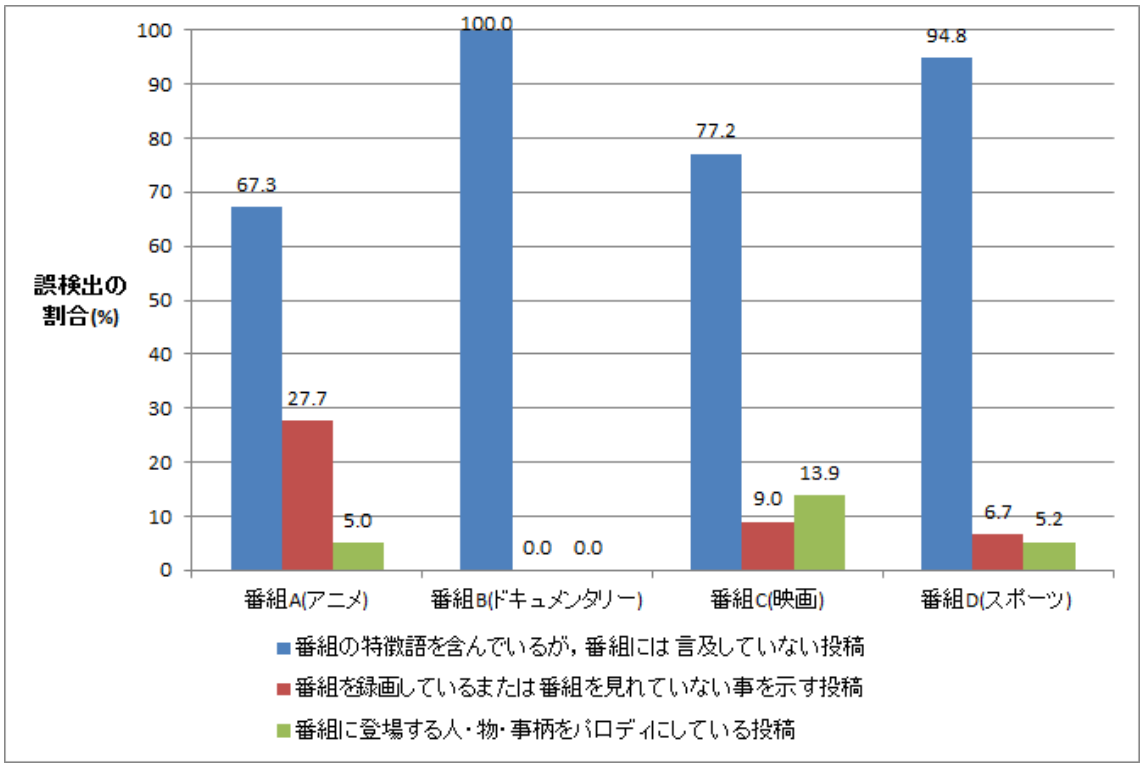

図 33　ジャンル別による誤検出の割合

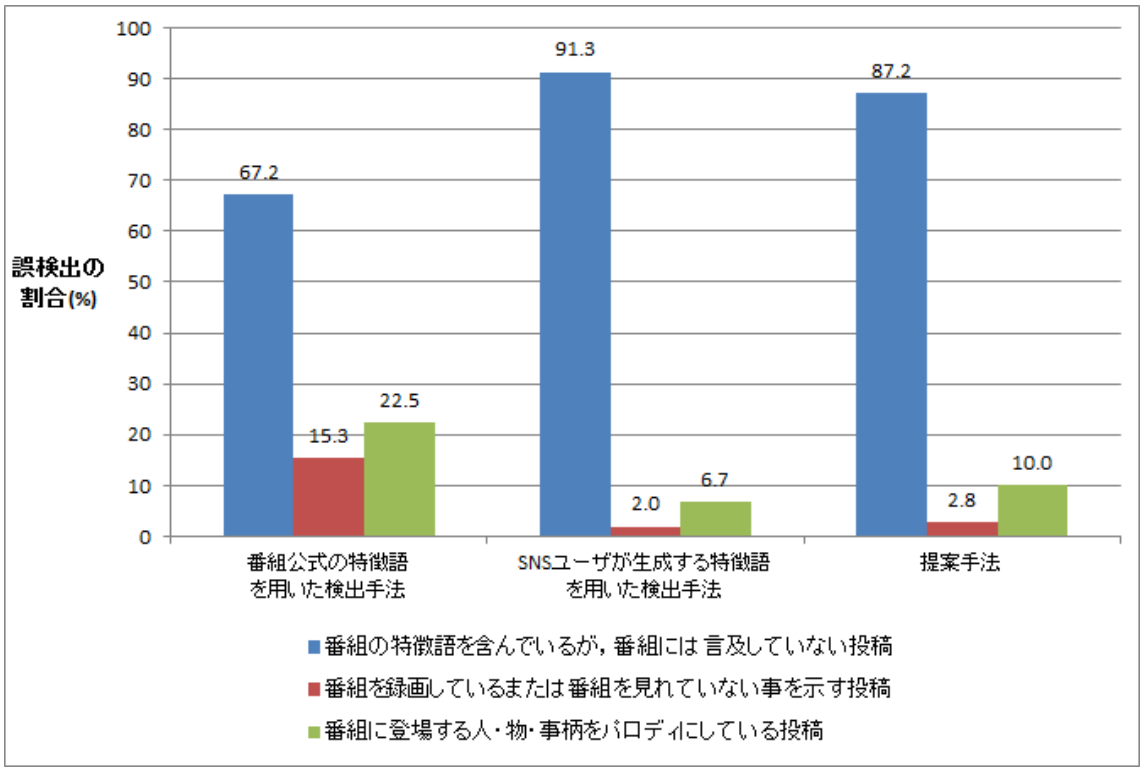

図 34　検出手法別による誤検出の割合

5. 4. 4 テレビ番組についての意見投稿を行ったユーザを検出できなかった原因についての考察

本項では，テレビ番組について意見投稿を行ったユーザを検出できなかった原因について考察を行う．まず，検出できなかったユーザの投稿は，以下に示すように 3 つに分類することができた．「2.意見の対象となった名詞が，番組の特徴語となる時間帯外であった投稿」とは，ちょうど図 35 に示すような投稿のことである．

1. 番組の放送時間帯を通して，番組の特徴語となり得なかった名詞を対象に意見している投稿
2. 意見の対象となった名詞が，番組の特徴語となる時間帯外であった投稿
3. 意見の対象となった名詞の綴りが誤っている投稿

意見投稿を検出できない原因の割合をテレビ番組のジャンルごとにまとめたものを図 36 に，検出手法ごとにまとめたものを図 37 に示す．

番組ジャンル別による意見投稿を検出できない原因は，「1.番組の放送時間帯を通して，番組の特徴語となり得なかった名詞を対象に意見している投稿」の割合が一番大きくなっている．番組の特徴語となり得ない名詞は，番組公式の特徴語による検出において，字幕テキストに現れない名詞や，あまりにも一般的な名詞であり，SNS ユーザが生成する特徴語による検出では，他のユーザがほとんど着目しなかった名詞だと考えられる．また，番組 D(スポーツ)において，「2.意見の対象となった名詞が，番組の特徴語となる時間帯外であった投稿」の割合が比較的高くなっている．番組 D(スポーツ)は，ロンドンオリンピック・男子サッカーについて放送された番組あった．サッカー試合の放送では，字幕テキストに選手名が現れなくても，テレビ画面で選手の行動を確認できるため，ユーザが着目した選手に対する批評を投稿が多く見られた．

手法別による意見投稿を検出できない原因でも，「1.番組の放送時間帯を通して，番組の特徴語となり得なかった名詞を対象に意見している投稿」の割合が一番大きくなっている．しかし，SNS ユーザが生成する特徴語による検出では，「2.意見の対象となった名詞が，番組の特徴語となる時間帯外であった投稿」の割合と同程度となった．これは，SNS ユーザが生成する特徴語による検出では，5 分毎に特徴語を作成しているため，番組に登場した人・物・事柄に少し遅れて投稿したユーザの意見を検出できないからだと考えられる．

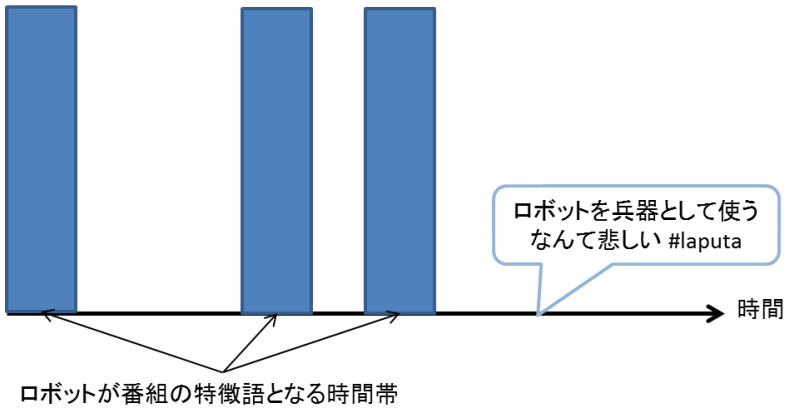

図 35　意見投稿を行ったユーザを検出できなかった原因の投稿例

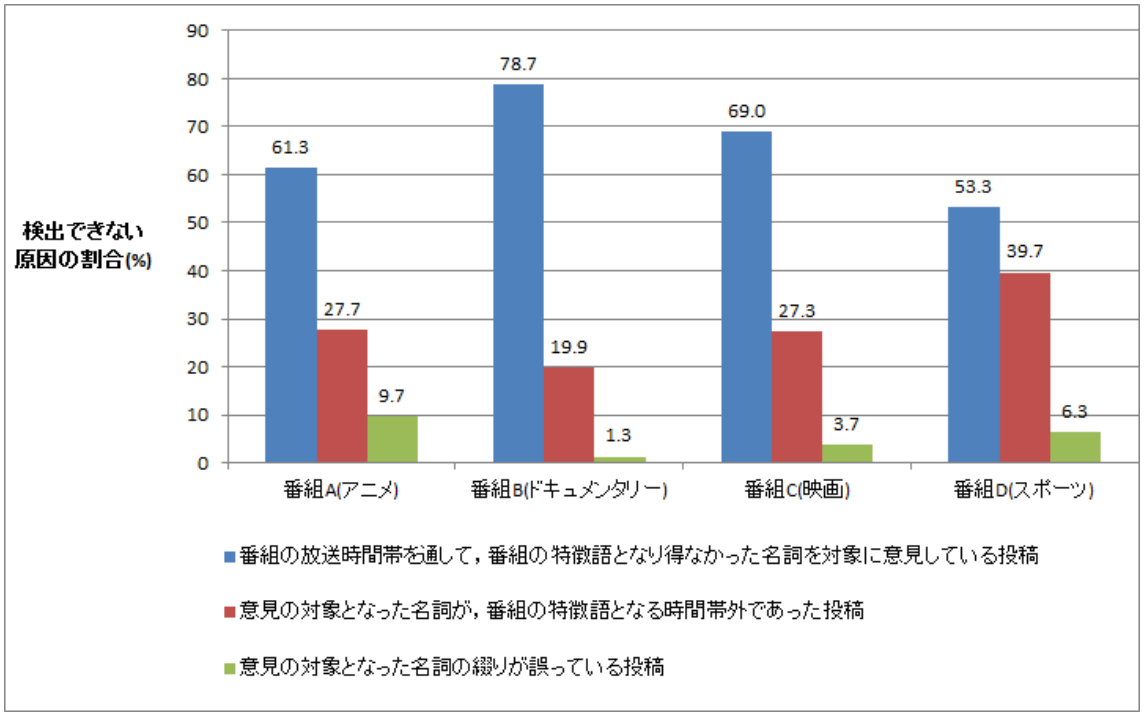

図 36　ジャンル別による意見投稿を検出できない原因の割合

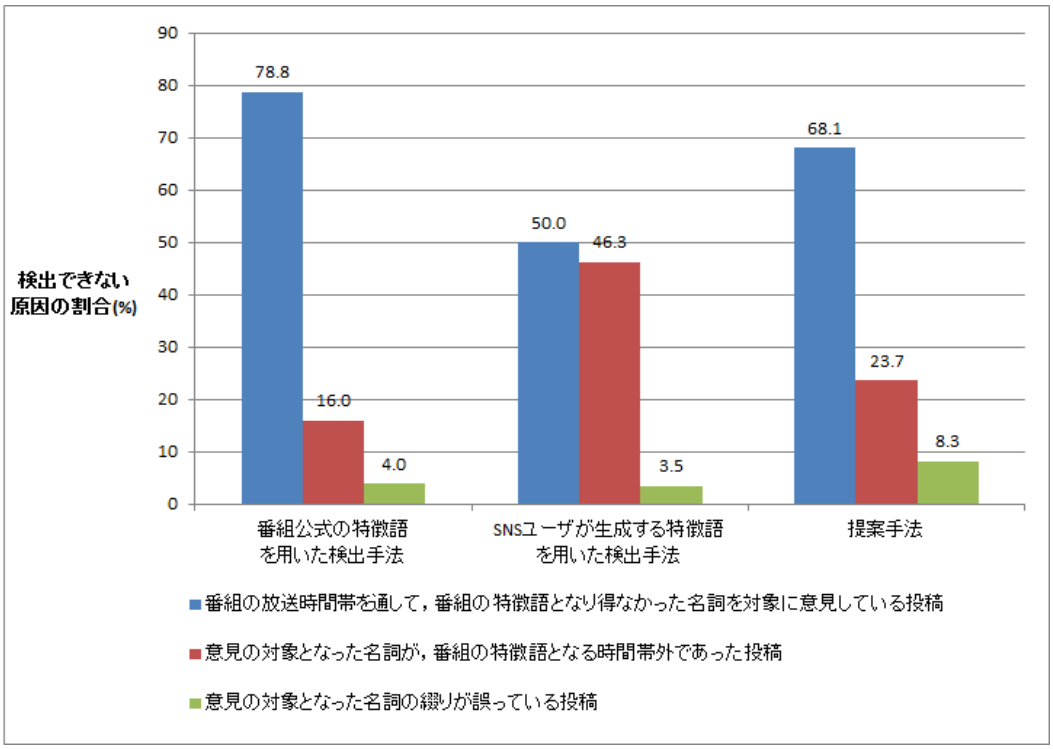

図 37　手法別による意見投稿を検出できない原因の割合

5. 4. 5 番組公式の特徴語群と SNS ユーザが生成した特徴語群の違いについての考察

本項では，番組公式の特徴語群でのみ検出できる投稿と，SNS ユーザが生成した特徴語群のみで検出できる投稿の違いについて考察を行う．

まず，番組公式の特徴語群でのみ検出できる投稿を抽出し，その投稿を検出するきっかけとなった番組の特徴語群を得た．この得られた特徴語群を，「番組公式の投稿検出特徴語群」とする．同様に，SNS ユーザが生成した特徴語群のみで検出を抽出し，その投稿を検出するきっかけとなった番組の特徴語群を得た．これを，「SNS ユーザが生成した投稿検出特徴語群」とする．次に，「番組公式の投稿検出特徴語群」および「SNS ユーザが生成した投稿検出特徴語群」に含まれる単語が，番組全体の字幕テキストに出現する回数を算出し，特徴語群ごとに出現回数の割合を求めた．番組 A(アニメ)，番組 B(ドキュメンタリー)，番組 C(映画)，番組 D(スポーツ)において求めた出現回数の割合を，それぞれ図 38～ 図 41 に示す．

4 番組に共通する傾向として，「SNS ユーザが生成した投稿検出特徴語群」には，「番組公式の投稿検出特徴語群」に見られない，字幕テキストに全く出現しない特徴語が含まれている．これは，字幕テキストに出現しない，SNS ユーザが着目した番組の特徴語によってのみ検出できる投稿があったことを示している．例えば，番組 A(アニメ)は，スタジオジブリが制作した「天空の城ラピュタ」の放送であり，字幕テキストには出現しない「ジブリ」という特徴語が得られた．また，番組 C(映画)は，「チャーリーとチョコレート工場」という洋画であり，出演俳優の吹き替えを担当した声優・宮野真守の苗字である「宮野」という特徴語が得られた．

さらに，「番組公式の投稿検出特徴語群」と「SNS ユーザが生成した投稿検出特徴語群」に重複して出現した単語が存在した．これは，特定の名詞が，番組公式の特徴語群(SNS ユーザが生成した特徴語群)になっていない時間帯において，SNS ユーザが生成した特徴語群(番組公式の特徴語群)になっており，その結果，検出できた投稿が存在していることを示している．よって，番組公式の特徴語群と SNS ユーザが生成した特徴語群は，特徴語となる名詞の種類だけでなく，時間の関係においても補完している部分があると言える．

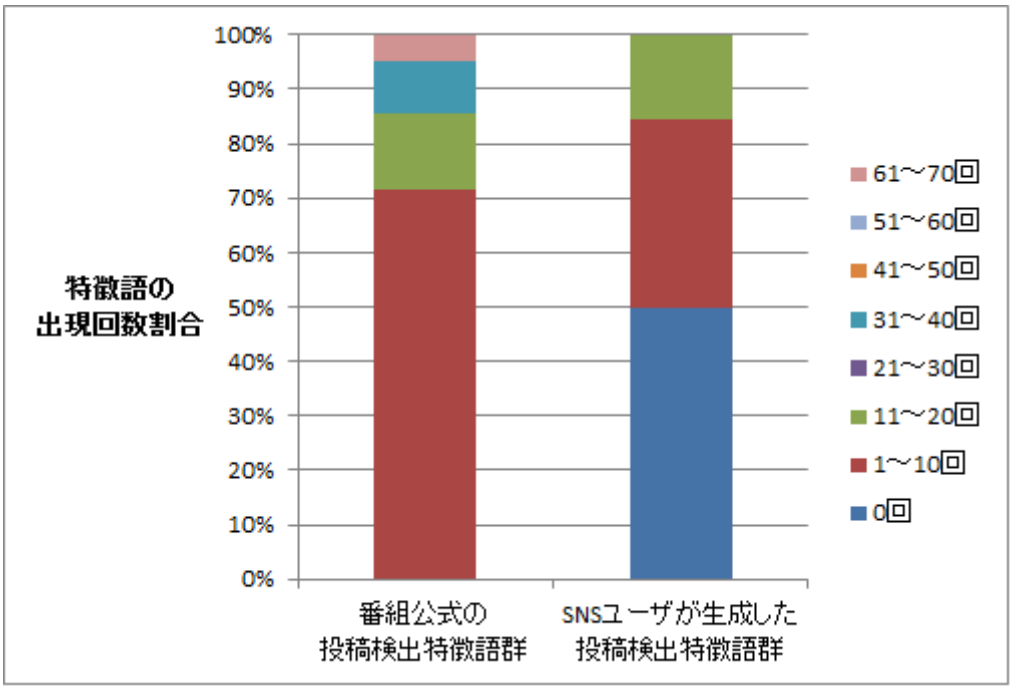

図 38　番組 A(アニメ)において，各投稿検出特徴語が字幕テキストに出現する回数の割合

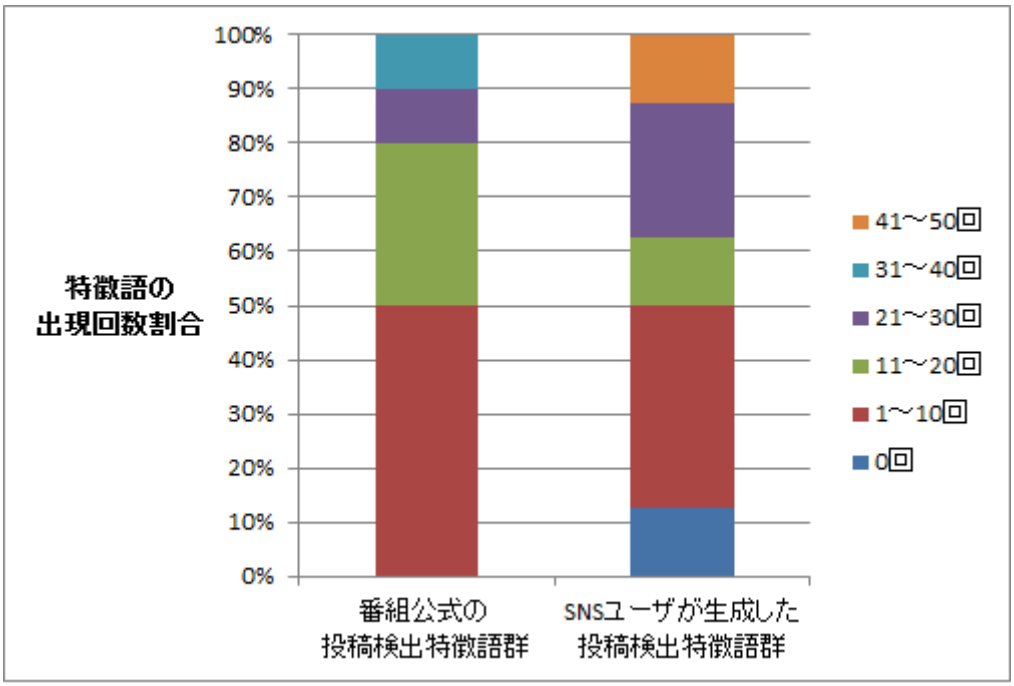

図 39　番組 B(ドキュメンタリー)において, 各投稿検出特徴語が字幕テキストに出現する回数の割合

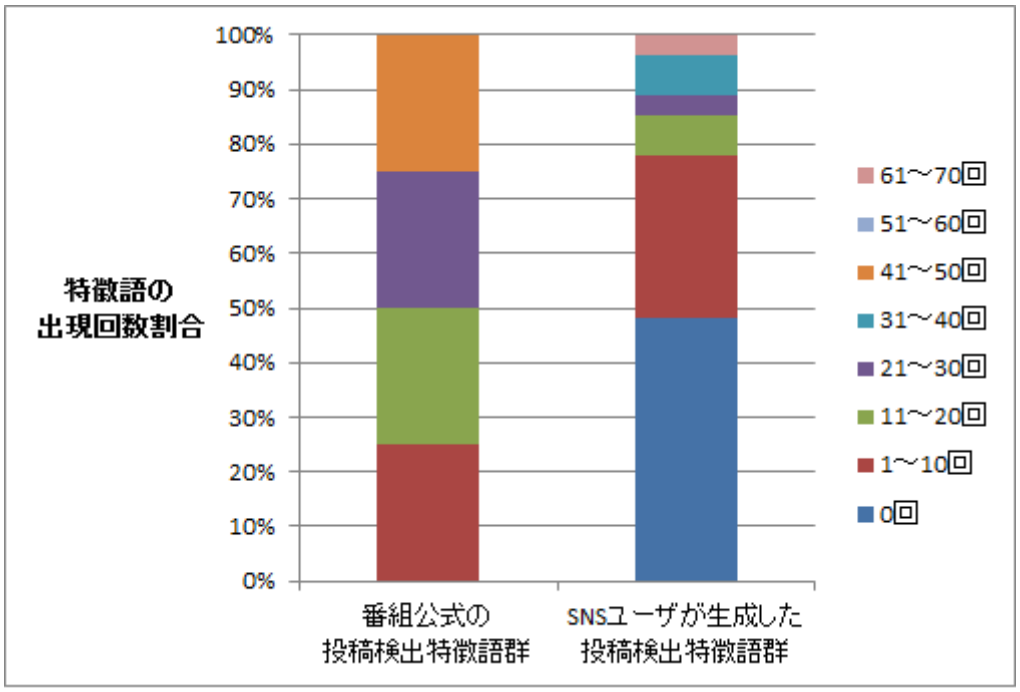

図 40　番組 C(映画)において，各投稿検出特徴語が字幕テキストに出現する回数の割合

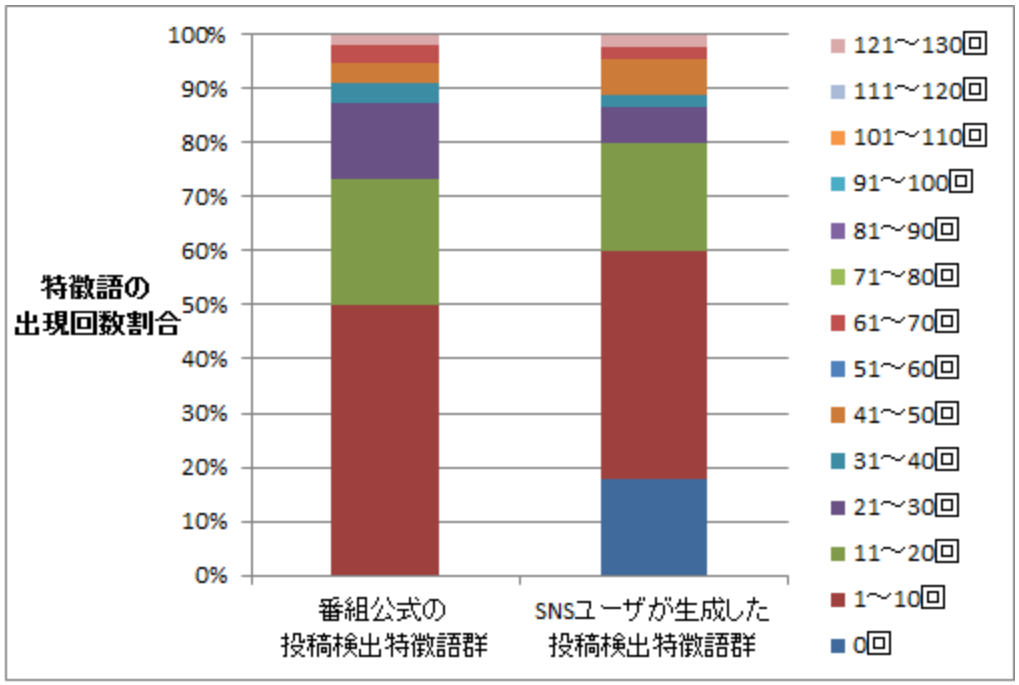

図 41　番組 D(スポーツ)において，各投稿検出特徴語が字幕テキストに出現する回数の割合

第6章　おわりに

本論文では，ソーシャルビューイングを行なっている Twitter ユーザから，意見を持ったユーザを発見する手法を提案した．実験の結果，提案手法は平均して 76%の適合率を保ちながら 68%の再現率で意見投稿を検出することができた．また，テレビ局やテレビ番組自体のハッシュタグを使用するユーザは一部であることがわかり，提案手法は，テレビ局やテレビ番組自体のハッシュタグを使用していないユーザの意見投稿を検出できていることを示せた．

今後の課題として，また，番組についての意見投稿を行ったユーザが，意見の対象となった番組中の人・物・事柄についてどのような意見を行なっているかを測ることを考えている．

謝　辞

本研究を行うにあたり，数々のご指導を頂いた山名早人教授に厚く御礼申し上げます．また，本研究を進めるに当たって，お忙しい中，数々のご指摘をしてくださった上田高徳先輩，Twitter のデータ収集に携わっていただいた浅井洋樹君に心から感謝いたします．に心から感謝いたします．

業　績

山本祐輔，及川孝徳，山名早人：”字幕テキストの利用によるマイクロブログからのテレビ番組に言及したメッセージ検出手法”，第 3 回データ工学と情報マネジメントに関するフォーラム，2011.

山本祐輔，浅井洋樹，上田高徳，秋岡明香，山名早人：” 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”，　第 5 回データ工学と情報マネジメントに関するフォーラム発表予定

参考文献

[1] 「Twitter、今年 6 月にユーザー5 億人超か―ブラジル急成長、ツイート数では日本語が依然英語に次いで 2 位」
http://jp.techcrunch.com/archives/20120730analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/
(2013/01/03 アクセス)

[2] 「Twitter の月間アクティブユーザーが 2 億人を突破」
http://www.itmedia.co.jp/news/articles/1212/20/news101.html
(2013 年/01/03 アクセス)

[3] 「SNS×TV 連携の現状と展望。Twitter/Facebook、mixi/LINE の取り組み」
http://av.watch.impress.co.jp/docs/news/20121018_566709.html
(2013 年/01/03 アクセス)

[4] 「ドラマ『SPEC～翔～』が「ソーシャルテレビ・アワード 2012」で大賞受賞」
http://dogatch.jp/news/tbs/9773
(2013 年/01/03 アクセス)

[5] ＴＶとＳＮＳの融合　つぶやき量、人気の指標に　（徳力基彦）　：日本経済新聞
http://www.nikkei.com/article/DGXNASGF3102T_R30C12A8H49A00/
(2013/01/03 アクセス)

[6] ニールセン、ツイッターを使ったテレビ視聴率調査サービス「Nielsen Twitter TV Rating」発表
http://markezine.jp/article/detail/16956
(2013/01/03 アクセス)

[7] ビデオリサーチ、Twitter をテレビ番組指標への活用に着手
http://markezine.jp/article/detail/16622
(2013/01/03 アクセス)

[8] Shoko Wakamiya, Ryong LEE, and Kazutoshi Sumiya: “Twitter-based TV Audience Behavior Estimation for Better TV Ratings”，第 3 回データ工学と情報マネジメントに関するフォーラム, 2011.

[9] Shoko Wakamiya, Ryong LEE, and Kazutoshi Sumiya: “Crowd-powered TV viewing rates: measuring relevancy between tweets and TV programs”, Proc of the 16th international conference on Database systems for advanced applications, pp.390-401, 2011

[10] Shoko Wakamiya, Ryong LEE, and Kazutoshi Sumiya: “Towards better TV viewing rates: exploiting crowd's media life logs over Twitter for TV rating”, Proc of the 5th International Conference on Ubiquitous Information Management and Communication, Article No39, 2011

[11] David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill: “Tweet the Debates: Understanding Community Annotation of Uncollected Sources”, Proc.of the 1st SIGMM workshop on Social media, pp.3-10, 2009.

[12] SayakaAkioka, Norikazu Kato, Yoichi Muraoka, and HayatoYamana: “Cross-media Impact on Twitter in Japan”, Proc.of the 2nd international workshop on Search and mininguser-generated contents, pp.111-118, 2010.

[13] 澤井　里枝, 有安　香子, 藤沢　寛, 金次　保明: “SNS を利用した協調フィルタリングによる番組推薦手法”，情報処理学会研究報告, Vol.2010-DBS-151 No.43, 2010.

[14] 加藤　慶一, 秋岡　明香, 村岡　洋一, 山名　早人: “ミニブログにおける注目語抽出手法の提案と注目語を用いたメディア間での話題追跡”，情報処理学会研究報告, Vol.2010-DBS-151 No.22, 2010.

[15] Kyoko Ariyasu, Hiroshi Fujisawa, and Yassuaki Kanatsugu: “Message Analysis Algorithms and their Application to Social TV”, Proc of the 9th international interactive conference on Interactive television, pp.1-10, 2011.

[16] Yuheng Hu, Ajita John, Dor’ee Duncan Seligmann, and Fei Wang: “What Were the Tweets About? Topical Associations between Public Events and Twitter Feeds”, Proc of the 6th International AAAI Conference on Weblogs and Social Media, pp.154-161, 2012.

[17] 「Twitter をテレビ番組制作に活用--TOKYO MX の BlogTV」
http://japan.cnet.com/news/media/20349032/
(2013/01/03 アクセス)

[18] 「朝まで生テレビ！が Twitter 連動　KDDI の分析エンジン活用、意見をリアルタイムに抽出 - ITmedia ニュース」
http://www.itmedia.co.jp/news/articles/1012/20/news101.html
(2013/01/03 アクセス)

[19] 「Twitter ユーザーの利用動向の実態は？『ソーシャルメディア白書 2012』掲載データを公開！」
http://markezine.jp/article/detail/15298
(2013/01/03 アクセス)

[20] 小林　尊志, 野田　雅文, 出口　大輔, 高橋　友和, 井手　一郎, 村瀬　洋: “Twitter における実況書き込み検出手法の検討”，映像情報メディア学会技術報告, Vol. 34, No.25, pp.129-130, 2010.

[21] Daniel Ramage, David Hall, Ramesh Nallati, and Christopher D. Manning: “Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora”, Proc of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.248-256, 2009.

[22] D. A. Shamma, and Y. Liu: “Social Interactive Television: Immersive Shared Experiences and Perspectives”, pp.273-288, Information Science Publishing, 2009

[23] 宮森　恒, 中村　聡史, 田中　克己: “番組実況チャットに基づく視聴者視点を利用した放送番組のビュー生成”, 電子情報通信学会第 16 回データ工学ワークショップ, 4B-i9, 2005

[24] 上原　宏, 吉田　健一: “インターネット上の対話文にもとづくドラマ番組の構造化注目状態グラフによる視聴者コミュニティの嗜好パターン認識”, 電子情報通信学会技術研究報告, Vol.104 No.369, pp.25-30, 2004

[25] emocon 友達とテレビ番組を楽しむアプリ
http://emocon.me/
(2013/1/17 アクセス)

[26] テレ Viewing Yahoo! JAPAN
http://promo.digitalhome.yahoo.co.jp/tvviewing/
(2013/1/17 アクセス)

[27] Gnip Providing Social Media Data for the Enterprise
http://gnip.com/
(2013/1/17 アクセス)

[28] Stanford Topic Modeling Toolbox
http://nlp.stanford.edu/software/tmt/tmt-0.4/
(2013/1/29 アクセス)