SNSからの1次情報自動抽出に向けた表層的言語情報の分析

福島裕斗* 桝井文人 Ptaszynski Michal 中島陽子
渡辺桂祐 河石良太郎 新田大征 佐藤亮弥
(北見工業大学)†

1 はじめに

Facebook や mixi のような知人とのコミュニケーションを目的とした SNS に比べ，ツイッターは情報発信に特化したサービスである．例えば,「リツイート」「ハッシュタグ」「拡散希望」といった独自機能を利用することで他の SNS よりも簡単に不特定多数のユーザーへの情報発信を可能としている．今日では，ツイッターは日常生活における重要な情報源となっており，個人の意思決定にも影響を与える存在であり，多くの研究が行われている [1, 2, 3, 4, 5, 6, 7]．例えば，ツイッターから観光情報を抽出する研究 [8] や，事実情報を判定してインフルエンザの流行予測を行う研究 [9] や，ツイッターのデマ情報を分析してデマの拡散を防止する研究 [10] などが行われている．また，東日本大震災において緊急時の情報発信手段としての有効性も指摘されており [11]，さらに，直近の参議院選挙では史上初めてインターネット選挙運動が解禁されるなど，SNS は社会の主要インフラになりつつあることが分かる [12].

一般に，物事の判断，緊急時の情報収集を行う際には情報の取捨選択が重要になる．

ツイッターには多様なジャンルと関連する個人の意見，話題とは無関係なツイートやデマ情報などが雑多に混在している．そのため，意思決定や状況判断をするためにツイートログの利用を想定すると，大量の玉石混淆なデータから有効なデータのみを自動抽出する手法が必要となる．このとき重要なのは,「情報の正確性」と「情報の均一性」を確保することである．

正確な情報を判別する基準として 1 次情報と 2 次情報という考え方[1] がある．1 次情報とは，発信者が直接見た，会った，直接聞いたというような，自らが仕入れた現場情報である．2 次情報とは，誰かが言っていた，書籍などに記述されていた，ＴＶで見た，インターネットに記述があったというような，第三者を介して得た間接的な情報である．

また，人は何かを評価する時や物事を決定する時，ほとんどの場合何らかの外的要因や情報によって心理的影響を受けることが指摘されている [13]．Kahneman[14] は「事実に基づく事」を阻害する要因として「認知バイアス」の存在を指摘している．認知バイアスには，不確かな事態に対する判断や曖昧な情報に基づいて予測や判断を行おうとする際に，初期値が判断に影響してしまう「アンカー効果」や，個人の先入観に基づいて他者を観察し，自分に都合のいい情報だけを集めて，それにより自己の先入観を補強する「確証バイアス」等がある[2]．これらの認知バイアスによってユーザーの意思決定に偏りが生じる恐れがある．

本研究では，正確性が高く均一性のある情報を得るために，東日本大震災ツイートログの分類結果を基にツイート分類ルールを定義する．さらに，そのルールに従って東日本大震災，総選挙ツイートログを分類した．前者は 1 次，2 次情報，後者は 1 次，1.5 次，2 次情報が必要であることが分かった．さらに，新たなルールに従って選挙のツイートログを分類し，分類ルールの有効性について検討した．

2 東日本大震災ツイートログの分類

本章では東日本大震災時のツイートの分類結果について述べる．

東日本大震災ビッグデータプロジェクト[3] で Twitter Japan 株式会社が提供していた 3 月 11 日から 1 週間のツイートを利用した．

地震発生前の 151 件を省いた後，無作為に 6,000 件を抽出し，1 次情報，2 次情報の定義を基に東日本大震災ビッグデータプロジェクトメンバー 6 人で分類を行った．

1 次情報とは，発信者が直接見た，会った，直接聞いたというような，自らが仕入れた現場情報である．2 次情報とは，誰かが言っていた，書籍などに記述されていた，ＴＶで見た，インターネットに記述があったというような，第三者を介して得た間接的な情報である．

1 次情報は 1,539 件 (26%)，2 次情報は 2,083 件 (36%)，その他 2,227 件 (38%) であった (Fig.1).

各次数に分類された情報の例を Table1 に示す．1 次情報には「寒い」「つらい」といったユーザ本人の状態をツイートしているものが非常に多く存在した．出現しているものには「私」や「僕」といった 1 人称を含むものが多く，その他に「なう」や体現止めを用いる表現も多数存在した．

2 次情報には 1 次情報よりも全体の情報量は多いが，被災地の外側に向けた情報がより多く存在しており，公式 RT[4] による被災地の周辺や外側のライフラインに関する情報のツイートが存在し,「○○とのこと」や「○○らし

*fukushima@ialab.cs.kitami-it.ac.jp
†北海道北見市公園町 165 番地

[1] http://www.ip-blog.net/2006/12/post_210.html

[2] http://moonwater.org/consul/04pointview/column/sub4_bias.htm

[3] https://sites.google.com/site/prj311/

[4] http://d.hatena.ne.jp/keyword/%A5%EA%A5%C4%A5%A4%A1%BC%A5%C8

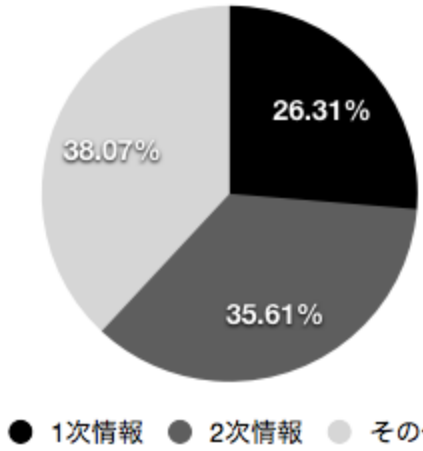

Fig. 1 東日本大震災ツイートにおける 1 次，2 次情報の割合

Table 1 東日本大震災ツイートの分類例

次数	例
1 次情報	冷蔵庫あいちゃって中全部落ちてきた よー しかし おなかすいた にゃー 停電きたぁ
2 次情報	RT @***:東急戦、世田谷線以外は前線再開。 らしい。心配。@RT***:東京来て一番でかい RT @***:RT してください!! 全国避難所一覧
その他	@*** (゜ロ゜) 笑 くぅぅぅ … 花見だと？

い」という表現が多数存在していた．

3 1 次，2 次情報分類基準の明確化

本章では東日本大震災ツイートログの分類結果を基に，定義した 1 次情報と 2 次情報の分類基準について記す．

1 次情報の内容について Table2，2 次情報の内容について Table3 に示す．

4 総選挙ツイートの分類

総選挙に関するツイートを分類するために，ハッシュタグクラウド[5] というハッシュタグ活用サービスサイトを利用した．このサイトでは，過去 1 週間分のツイートをハッシュタグごとに保存，提供している．そこで,「＃総選挙」で検索し，2012 年 12 月 3 日から現在までのツイートを分析用データとして 1 日ごとに取得した．

そのうち，2012 年 12 月 3 日 ~4 日 (第 46 回衆議院議員総選挙公示日) の 1,503 件のツイートを対象に分析した．「＃総選挙」によって取得したツイートは通常のツイートの平均文字列長[6] に比べ 30 文字程長かった．また，断定表現が多く，リプライ数は平均[7] の 23%に比べ，5%と非常に少なかった．

1 次情報は 1,503 件中 1,317 件 (88%)，2 次情報は 186 件 (12%) であった (Fig.2)．

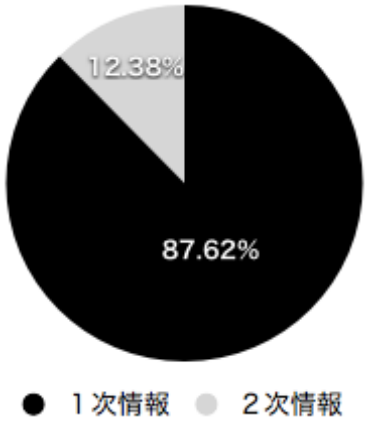

Fig. 2 総選挙ツイートログにおける 1 次，2 次情報の割合

また，非公式 RT82 件のうち，1 次情報として分類されたものは 57 件,「＃ RT」64 件のうち，1 次情報として分類されたものは 52 件であった．非公式 RT には政党の紹介ツイートに対して応援のコメントを書いているものや，憲法改正のツイートに対して自分の意見を述べているものがあった.「＃ RT」には，選挙支援サイトへのリンクを貼っているものがあった．

2 次情報には公式 RT で選挙ポスターの写真を貼付けているものや,「○○で選挙演説してるらしい」というようなツイートがあった．

さらに，全ツイート 1,503 件中 997 件が URL 情報を含むツイートで，投票の参考となりうる情報が多く存在した．例えば，ツイート中には地区ごとの選出議員一覧へのリンクが 197 件あったが，これらは 1 次情報を含むものが非常に多く，政党にとってポジティブな表現やネガティブな表現を含むものもあった．

有用な情報を自動抽出するためには，まず正確性の高い 1 次情報の表層的な表現に注目すべきである．1 次情報 1,317 件から無作為に抽出した 1,000 件の 1 次情報をツイートの内容から政党にとってポジティブ，ネガティブ，ニュートラルの 3 つに分類した．また，客観的な意見はニュートラルとして分類した．

分類の結果，上記の 1,000 件中，ポジティブなツイートは 68 件 (7%)，ニュートラルなツイートは 771 件 (77%)，ネガティブなツイートは 161 件 (16%) であった (Fig.3)．

ポジティブなツイートには,「○○党応援してます!!」など特定の政党や候補者の良いイメージを与えるもの，ネガティブなツイートには,「○○党は絶対に投票しない」など特定の政党や候補者の悪いイメージを与えるもの，

[5] http://hashtagcloud.net/
[6] http://teapipin.blog10.fc2.com/blog-entry-294.html
[7] http://b.hatena.ne.jp/entry/www.tommyjp.com/2010/10/7123rt6.html

Table 2 総選挙ツイートログ分類時の 1 次情報の定義

分類基準	例
直接「見た」「聞いた」「行動した」等の 自分で確認することのできた事実内容のツイート	青森県内の衆院選の立候補予定者動画を撮影しました
断定的な表現 (「〜だ」「〜である」など) を含むツイート	
非公式 RT で自分の意見を書いているもの 非公式 RT：他のユーザーのツイートを引用し さらに自分の書いたことも一緒に載せるもの	支持組織の自治労の不始末を税金突っ込んでリカバリしただけで自慢できる実績ではないよね。 RT @
「拡散希望」でリツイートされていない元のツイート 「拡散希望」: 宣伝目的のツイートを他のユーザーに リツイートしてもらうときに付けられるキーワード	【拡散希望】福岡 10 区の全候補者の政策を 動画でチェックできます。
「＃ RT」(拡散希望と同義) で内容に伝聞推定 (「〜らしい」「〜みたい」など) が含まれていないもの	選挙情報サイト「エレクトペディア」サイトの周知に ご協力お願い致します。 #RT
「なう (現在自分が行っているという意味の用語)」 を含むもの	明日からの選挙に向けて 学習なう。(と言いつつネットなう

Table 3 総選挙ツイートログ分類時の 2 次情報の定義

分類基準
ニュースサイトなどの引用ツイート (URL 情報やツイート内容「○○ニュース」等の表記から判断)
公式 RT：他のユーザーのツイートを引用形式で自分のアカウントから発信することであり 自分が興味を持った誰かのツイートを手軽にフォロワーへと流すことができる
「見た」「聞いた」「らしい」等の伝聞推定の表現を含むもの
非公式 RT で自分の意見を書いていないもの

Table 4 再定義したツイート分類ルール

ツイート	1 次	1.5 次	2 次	例
事実情報	○			ネット選挙解禁後、歴史的な最初の選挙
行動の記述	○			選挙行ってきた
断定表現	○			山本太郎さん当確
インタビューの内容	○			政治家を悪者にしてもなにもはじまらない/菅原琢氏インタビュー
政策	○			田宮かいちの政策 1，業界団体ではなく、あなたの声を国政に
意思表示		○		選挙行く !!
感情表現		○		山本太郎 おめでとう!!嬉しいです
意見		○		山本太郎、もっと簡潔に喋ってくれ
呼びかけ		○		棄権しないでしっかり投票しましょう！
リンクへの誘導		○		議員がどのように考えているか。 ここでチェック！
公式リツイート			○	
TV で見た (実況含む)			○	今回も安定の池上さんで開票速報！
伝聞推定表現			○	市議会議員選挙妨害しているこの鹿、どうも野党支持らしい
転載元の書かれたもの			○	東京選挙区敗北なら辞任の意向 - 朝日新聞デジタル
著名人の言葉の引用のみ			○	大きなる事を謀るには、輔有るには如かず。by 中臣鎌子

ニュートラルなツイートには，選挙区ごとの候補者のリストなどがあった．

5 分類ルールの再定義

東日本大震災のツイートログ，総選挙ツイートログの分析結果に基づき，ツイッターに対する分類基準を再定

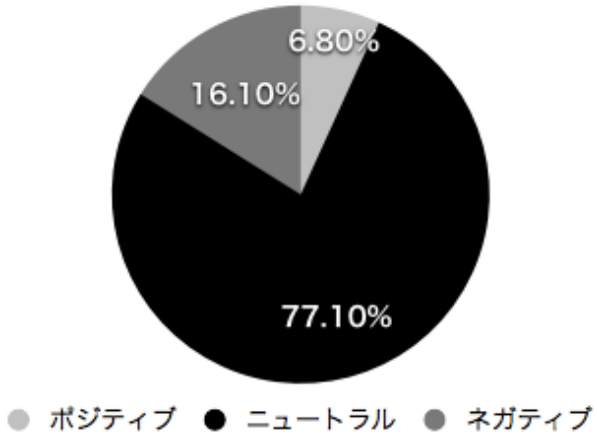

Fig. 3 総選挙ツイートログにおける極性による分類

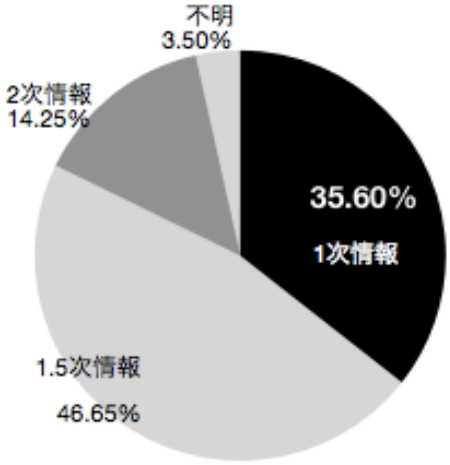

Fig. 4 選挙ツイートログにおける分類結果の内訳

義した．事実情報が重要な東日本大震災ツイートの分析結果を基に定義した分類基準を投票判断にする総選挙ツイートログに適用するには，必要な情報に違いがある．投票判断に資する情報は正確な情報だけでなく，参考となる情報も重要である．

こうした情報には意見や意思など，震災時のツイートではノイズとされる情報も含まれる．震災時にはノイズとなりうる情報でも，選挙においては個人の意見として参考になる情報である．これらの情報は混在させて提示させるのではなく，明確に判別して提示すべきである．そこで，1 次情報，2 次情報に属さない中間的な情報として 1.5 次情報を定義した (Table4)．

ツイートには情報の混在が多い．このコンフリクトを解消するために，(1) ツイート内で次数が輻輳した場合は，より次数の低い情報を優先する，(2)1.5 次情報と 2 次情報のみが現れた場合は 2 次情報を優先する (例:〜だと思う．○○ニュース) というヒューリスティクスを設定した．

6 選挙ツイートの分類

ハッシュタグクラウドを利用し，選挙に関するツイートとして第 23 回参議院議員通常選挙の公示日 (2013 年 7 月 4 日) から投開票日 (7 月 21 日) のツイート 22,176 件を取得した．

取得したデータから明確に 2 次情報と判断できる公式リツイート (93 件) を予め抽出し，残りの 22,083 件から 2,000 件を無作為抽出して調査を行った．選挙ツイートログの平均文字列長は 45 字で，通常のツイートの平均文字列長[8] より 15 字程長かった．これは通常のツイートより情報量が多い事を示唆している．1 次情報は 711 件 (36%)，1.5 次情報は 933 件 (47%)，2 次情報は 286 件 (14%) であった (Fig.4)．

さらに，どの次数にも分類できないツイートが 70 件 (3%) 存在した．分類できないものには，ハッシュタグのみのツイートや，地名のみのツイート，あいさつのみのツイートなどがあった．また，1 次情報に分類したツイートの約 90%が 1.5 次情報とともに出現した．1 次情報に分類された非公式リツイートは 29 件，1.5 次情報に分類されたのは 135 件であった．

有用な情報を自動抽出するために，正確性の高い 1 次情報の表層的な表現に注目したところ，1 次情報中でポジティブ，ネガティブな印象を与えるものは存在せず，全てがニュートラルなツイートであることが確認された．

7 考察

東日本大震災ツイートにおいて，1 次情報には正確性の高い情報が含まれていたが，被災地においてノイズとなる情報が含まれていた．分析者がそれぞれの主観により分類したため，分析者ごとに分類基準に差異が生じる．そのため，分類には明確な基準を設け，それに沿った分類を行う必要がある．

総選挙ツイートログの非公式 RT における 1 次情報の数を調べると 82 件中 57 件存在していた．分析前には非公式 RT のほぼすべてが 1 次情報に分類されると予想していたが，結果を見ると非公式 RT の 30%が 2 次情報に分類され，予想と大きな違いが現れた．これは，公式 RT にはしたくないが拡散したいというツイートが出てきてしまったためであると思われる．

また，URL 情報を含むツイートには投票の参考になるツイートが多く存在していた．これは，情報が正確かどうかの判断材料として有用である．

1 次情報における意見情報に注目すると，ニュートラルな表現のツイートが 77%になった．ニュートラルなツイートは，客観的な目線からの中立的意見であり，有権者が投票する際の判断材料になる．つまり，ニュートラルなツイートを自動抽出できれば，有用な情報の抽出ができるのではないかと考える．また，ポジティブやネガティブな意見のツイートは，個人の偏った考えを押し付

[8] http://teapipin.blog10.fc2.com/blog-entry-294.html

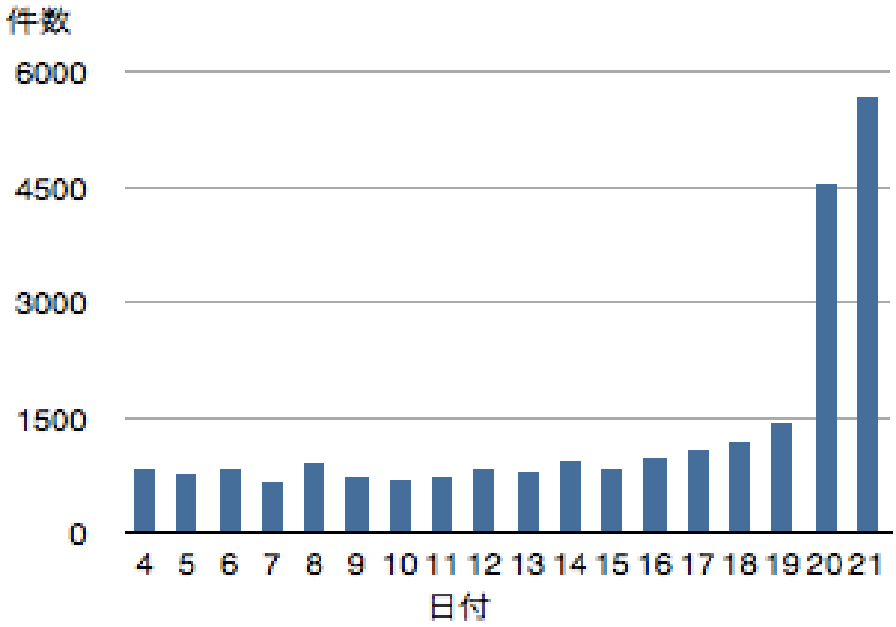

Fig. 5 1 日ごとのツイート件数

ける形になるため，選挙ツイートにおいて有用な情報にはなり得ないと考えられる．しかし，政党ごとのポジティブ，ネガティブの割合は参考情報となる可能性があるため，より厳密な判別基準が必要である．

12 月 3 日 (公示日前日) は 623 件，12 月 16 日 (投開票日) には 13,093 件のツイートがあった．投開票日に近づくにつれツイート数が増加し，16，17 日にピークとなり，その後減少している．投開票日に大量のツイート数があるということは，選挙の結果に注目するユーザーがとても多かったということである．そのため，公示日前後，投開票日前後ではツイートの種類に異なる傾向が出る可能性がある．

2 つの分析結果を基に再定義したルールで分類した選挙ツイートログでは 1.5 次情報を追加することで 1 次情報をより正確に抽出できたが，「あいさつ」など，一部のツイートは今回の基準でも分類することができなかった．これらの扱いについては今後更に検討していく必要がある．

結果を見ると 1.5 次情報が 47%と非常に多く，ツイートの文字列長も平均より長いことから，他人の意見をリツイートするよりも自分の意見を述べようという政治への関心が伺える．ところが，実際の投票率は戦後 3 番目の低さとなっている[9] ことからツイートにおける意思表示行動が必ずしも投票行動に反映されていないことが示唆される．

Fig.5 から分かる通り，ツイート件数は投開票日に近づくにつれて増加しているため，選挙自体に関心が無いということはない．これは，ツイッターユーザーと投票に参加した人の年齢層に関係すると思われる．ツイッターユーザーの割合は年齢が上がるごとに減少し[10]，投票率は年齢が上がるごとに増加している．このことから 2013 年 4 月 19 日にインターネット選挙運動が解禁されたが，投票まで確実に結びついていないと考えられ，ツイート数の増減と投票率の増減に関係はないことが示唆される．また，非公式リツイートが 2 次情報に分類された件数が 0 であったことから，非公式リツイートは 2 次情報になり得ないと結論付けられる．

1 次情報に分類された意見情報を見ると，分類された全てのツイートがニュートラルなものであった．この結果は，認知バイアスを考慮するために考案した 1.5 次情報によって，1 次情報に紛れ込むノイズを除去できたものといえる．ツイートの中には「ネット選挙解禁後、歴史的な最初の選挙。みんなで投票にいこう。」のように複数の次数を含むものがあるが，今回の調査では区別していない．このことは，例えば 1 次情報と 1.5 次情報が共存している状態で存在していることを意味し，1.5 次情報として記述されているポジティブ/ネガティブな意見が 1 次情報ツイートに含まれていることになる．これは，ユーザーに提示した時に認知バイアスを発生させる要因となり得るため，1 次情報に分類したツイートの内容をさらに分類する，あるいはあらかじめツイートを文単位に分割しておく対策が必要となる．

8 おわりに

意思決定・状況判断をするための基礎データとして，SNS を用いる状況を想定し，1 次情報自動抽出に向けて震災データを分析し，その結果を基に分類定義を行った．

[9] http://sankei.jp.msn.com/politics/news/130722/elc13072202420023-n1.htm

[10] http://web-tan.forum.impressrd.jp/e/2012/05/11/12694

また，定義を基に総選挙ツイートログを分類した結果，認知バイアスに影響のある情報が 1 次情報に存在し，1 次情報の中からさらに事実情報と意見情報を判別する必要がある事が分かった．

そこで，東日本大震災ツイートと総選挙ツイートの分析結果を基に分類ルールを再定義し，選挙ツイートログを対象として分析を行った．ハッシュタグクラウドから 7 月 4 日 ~21 日の「#選挙」ツイートを取得，2,000 件を無作為抽出し分析した．その結果，認知バイアスを考慮して再定義した分類ルールの有効性を確認できた．1 次情報における意見情報を調査した結果，すべてのツイートがニュートラルなものとなったが，一緒に書かれている 1.5 次情報に認知バイアスがかかるため，何らかの処理をする必要がある．今後，1 次情報自動抽出システムを構築，人手で分類した結果と比較し，システムの性能評価を行う予定である．

謝辞

本研究の一部は，科学研究補助金 (基盤研究 (C) 20500833) の助成を受けている．

参考文献

[1] 田中淳史，田島敬史：“ twitter のツイートに関する分類手法の提案 ”，第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM 2010)，A5-4，2010-3

[2] 風間一洋，今田美幸，柏木啓一郎：“ ツイッターの情報伝播ネットワークの分析”，第 24 回人工知能学会全国大会，2010

[3] 岩木祐輔，ヤトフトアダム，田中克己：“マイクロブログにおける有用な記事の発見支援”，第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009)，March 2009

[4] Long Jiang，Mo Yu，Ming Zhou，Xiaohua Liu，Tiejun Zhao：“ Target-dependent Twitter Sentiment Classification”，ACL 2011

[5] 神嶌敏弘，“2 協調フィルタリングの課題：プライバシー，サクラ攻撃，評価値のゆらぎ”，情報処理，vol.48，no.9，pp.966-971，2007

[6] 藤坂達也，李龍，角谷和俊：“地域イベント発見および特性検証のための実空間マイクロブログを用いたユーザ移動パターン分析システム”，情報処理学会全国大会講演論文集 第 72 回，平成 22 年 (1)，”1-845”-”1-846”，2010-03-08

[7] 梅島彩奈，宮部真衣，荒牧英治，灘本明代：“災害時 Twitter におけるデマとデマ訂正 RT の傾向”，情報処理学会研究報告，データベース・システム研究会報告，2011-DBS-152(4)，1-6，2011-07-26

[8] 桑野孝光，三田村保，渡辺功，鈴木康広，大堀隆文：“ Twitter を利用した観光情報の調査分析 ”，観光情報学会誌，Vol.8，No.1，pp.27-38(2012)

[9] 荒牧英治，増川佐知子，森田瑞樹：“ 事実性判定を用いたインフルエンザ流行予測”，情報処理学会研究報告．SLP，音声言語情報処理 2011-SLP-86(1)，1-8，2011-05-09

[10] 梅島彩奈，宮部真衣，荒牧英治，灘本明代：“災害時 Twitter におけるデマとデマ訂正 RT の傾向”，情報処理学会研究報告，データベース・システム研究会報告，2011-DBS-152(4), 1-6， 2011-07-26

[11] 立入勝義：“検証 東日本大震災 そのときソーシャルメディアは何を伝えたか?”，ディス カヴァー・トゥエンティワン (2011)

[12] 宮部真衣，荒牧英治，三浦麻子：“ 東日本大震災における Twitter の利用傾向の分析”，情報処理学会研究報告．GN，[グループウェアとネットワークサービス] 2011-GN-81(17)，1-7，2011-09-08

[13] 小林卓弥，大島裕明，小山聡，田中克己：“ レビュアーのプロファイルと地域性に起因するバイアス補正に基づくレビュー情報の信憑性向上”，第 19 回データ工学ワークショップ (DEWS2008) 論文集

[14] Kahneman，D；Tversky，A：“Subjective probability：A judgment of representativeness”，Cognitive Psychology 3：430-454，1972