3O1-11in

表層的言語情報から読みとれる直接性に着目したツイートログの分類

Classification of Tweet Logs Based on Directness Derived from Surface Expressions

福島 裕斗[*1] Taisei Nitta　桝井 文人[*1] Fumito Masui　プタシンスキ ミハウ[*1] Michal Ptaszynski　中島 陽子[*1] Yoko Nakajima　渡辺 桂祐[*1] Keisuke Watanabe
河石 良太郎[*1] Ryoutarou Kawaishi　新田 大征[*1] Taisei Nitta　佐藤 亮弥[*1] Ryoya Satou

[*1]北見工業大学 情報システム工学科
Department of Computer Science, Kitami Institute of Technology

In this paper we present our study in classification of tweet logs appearing during two kinds of events - disasters and elections. We analyze the credibility of information included in tweets according to two criteria differing in directness of the expressed information. In particular, we classify tweets into Primary and Secondary Information. We also perform sentiment analysis of tweets to verify the aptness of the proposed criteria.

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．今日のツイッターは日常生活における重要な情報ソースとなっており，個人の意志決定に影響を与える存在である．

また，ツイッターを利用した研究も活発に行われている [岩木祐輔 2009][風間一洋 2010][福島裕斗 2013]．例えば，ツイッターから観光情報を抽出する研究 [桑野孝光 2012] 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 [梅島彩奈 2011] 等がある．さらに，東日本大震災において緊急時の情報発信手段としてツイッターの有効性も指摘されており [立入勝義 2011][宮部真衣 2011]，2013 年 7 月に行われた参議院議員選挙では史上初めてのインターネット選挙運動が解禁され，議員が広報活動にツイッターを利用する等，今やツイッターは社会の主要インフラとなっている．

一般に，意志決定・状況判断に伴う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，ユーザーの知りたいトピックのツイートと，それに全く関係のないツイートが雑多に混在している．そのため，意志決定や状況判断をするための情報ソースとしてツイートログを利用する状況を想定すると，大量の玉石混淆なデータから有効なデータを自動抽出し，その情報を選別する必要がある．このような，情報を重要度や緊急度に応じて分類，選別することを情報トリアージ [Catherine C. Marshall 1997][Sofus A. Macskassy 2001][Macskassy 2011] と呼ぶ．情報トリアージとは，時間的・資源的制約があって任務や課題のすべてを実施・完了できないとき，一定の基準に従って着手の優先/非優先を判断することである[*1]．

情報トリアージを行う際に重要となるのは,「情報の正確性」と「情報の均一性」を確保することである.「情報の正確性」を判別する基準として 1 次情報・2 次情報という考え方[*2]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，2 次情報とは,「誰かが言っていた」「書籍等に記述されていた」「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．

「情報の均一性」を判別する基準としては，認知バイアスの影響に考慮する．Kahneman[Kahneman.D,Tversky.A 1972] 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「アンカー効果」[*3]や,「確証バイアス」[*4]等がある[*5]．このような認知バイアスの影響によってユーザーの意志決定に偏りが生じる恐れがある．

そこで本研究では，情報トリアージの定義にのっとり，正確性が高く均一性のある情報を自動抽出する手法について検討する．第 2 章では，緊急を要する決断に伴う場合を想定した東日本大震災ツイートログ (3 月 11 日から 1 週間のツイート，約 1 億 7000 万件)[*6]と，複雑な意志決定プロセスを伴う場合を想定した選挙ツイートログ (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)[*7]を人手によって分析し，その結果を基にツイート分類基準を作成する．第 3 章では，基準に従って震災ツイートログ，総選挙ツイートログを分類する．第 4 章では，評価実験の結果から考察を行い，分類基準の有効性について検討する．

2. 分類基準の作成

本章では，緊急を要する決断に伴う場合と，複雑な意志決定プロセスを伴う場合について，ツイートログの分類結果から 2 つの分類基準を作成する．以下，各基準について説明する．

連絡先: 福島裕斗，北見工業大学 情報システム工学科，北海道北見市公園町 165

∗1 http://www.itmedia.co.jp/im/articles/0612/12/news117.html

∗2 http://www.ip-blog.net/2006/12/post_210.html

∗3 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すること

∗4 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こと

∗5 http://moonwater.org/consul/04pointview/column/sub4_bias.htm

∗6 以降，震災ツイートログとする

∗7 以降，選挙ツイートログとする

2.1 緊急を要する決断に伴う基準:(基準 1)

緊急を要する決断を行う場合，情報ソースに最も求められるのは正確性である．情報の正確性に着目したものとして 1 章で述べた 1 次情報・2 次情報という考え方がある．

1 次情報とは，発信者が直接見たり，会ったり，聞いたりすることで自らが仕入れた現場情報である．それに対して 2 次情報とは，他人から聞いたり，書籍等に記述されていたり，ＴＶで見たり，インターネットの記述を見たなどの，第三者を介して得た間接的情報である．これらの定義を用いることで緊急時に求められる情報の正確性に考慮した分類ができると考えた．

以上の定義と，震災ツイートログから無作為抽出した 6,000 件の分析結果を基に，ツイッターにおける 1 次情報・2 次情報の分類基準を作成した (表 1, 表 2)．

2.2 複雑な意志決定プロセスに伴う基準:(基準 2)

投票判断等の複雑な意志決定プロセスを伴う場合は，緊急を要する決断を伴う場合と比べ，必要な情報に違いがある．複雑な意志決定プロセス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．また，認知バイアスの影響を考慮すると，参考となる情報は 1 次情報等と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，意思や意見を含む参考となる情報を 1.5 次情報と定義した．以上の定義と，選挙ツイートログ 2,000 件の分析結果を基に分類基準を作成した (表 3)．

また，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，

- 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．
- 1.5 次情報と 2 次情報のみが現れた場合は 2 次情報 (例: ～だと思う．○○ニュース) を優先する．

というヒューリスティクスを設定した．

3. 評価実験

2 章で述べた基準に従って総選挙ツイートログ，震災ツイートログを分類し，分類基準の有効性を検証する．評価尺度として分類結果の冗長性[*8]を考える．以下，実験結果について述べる．

3.1 総選挙ツイートログの分類

本節では「#総選挙」によって収集したツイートログの分類結果について述べる．

3.1.1 実験環境

総選挙に関するツイートを分類するために，ハッシュタグを利用した．そこで,「＃総選挙」の 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[*9]．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類した．

総選挙ツイートログは通常のツイートの平均文字列長[*10]43 文字に比べ 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[*11]の 23%に比べ，5%と非常に少なかった．

3.1.2 分類結果

基準 1 による分類の結果，1 次情報は 1,317 件 (87.62%)，2 次情報は 186 件 (12.38%) であった (図 1)．

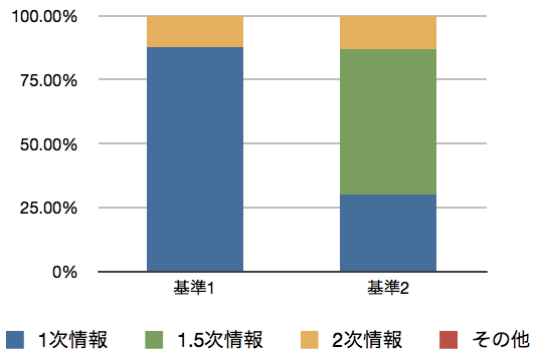

図 1: 総選挙ツイートログの分類

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多かった.「＃ 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．

基準 2 による分類の結果，1 次情報は 449 件 (30.09%)，1.5 次情報は 849 件 (56.90%)，2 次情報は 194 件 (13.00%) であった．

さらに，どの次数にも分類できないツイートが 11 件 (0.70%) 存在した．分類不能には選挙ツイートログに出現した挨拶のみのツイートはなく,「#総選挙 RT @○○:維新・自民の後ろには，橋下徹や … 」といった非公式 RT の形でハッシュタグを付与しているものが見られた．

3.2 震災ツイートログの分類

本節では，震災ツイートログの分類結果について述べる．

3.2.1 実験環境

対象データとして，震災ツイートからテストデータに用いたツイート 6,000 件を除き，600 件を無作為抽出した．

このデータは「#総選挙」のようにトピックを絞って集めたツイートと違い，地震発生後全てのツイートを対象としている．そのため，ツイートの平均文字列長 (43 文字) もリプライ数 (23%) も平均と近い値となった．

3.2.2 分類結果

基準 1 による分類の結果，1 次情報は 183 件 (30.50%)，2 次情報は 148 件 (24.67%)，その他は 269 件 (44.83%) であった (図 2)．その他には「@○○ん j ご」のように意図を理解できないものや,「Yahoo 募金」といった単語のみのツイートが存在した．その他に分類された割合が総選挙ツイートの分類結果に比べ非常に多く,「津波大丈夫かなぁ … 」といった曖昧な表現のツイートが目立った．

基準 2 による分類の結果，1 次情報は 141 件 (23.50%)，1.5 次情報は 210 件 (35.00%)，2 次情報は 148 件 (24.67%)，その他は 101 件 (16.83%) であった．

*8 本評価における冗長性の高さとは，分類結果に対して必要以上の情報が抽出されていることを指す．
*9 以降，総選挙ツイートログとする
*10 http://teapipin.blog10.fc2.com/blog-entry-294.html
*11 http://b.hatena.ne.jp/entry/www.tommyjp.com/2010/10/7123rt6.html

表 1: 1 次情報の分類基準 (基準 1)

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情報か？
断定的な表現 (「〜だ」「〜である」など) を含むか？
非公式 RT で自分の意見を書いているか？ (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)
「拡散希望」でリツイートされていない元のツイートか？ (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)
「＃ RT」(拡散希望と同義) で内容に伝聞推定 (「〜らしい」「〜みたい」など) が含まれていないか？
「なう (現在自分が行っているという意味の用語)」が含まれているか？

表 2: 2 次情報の分類基準 (基準 1)

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

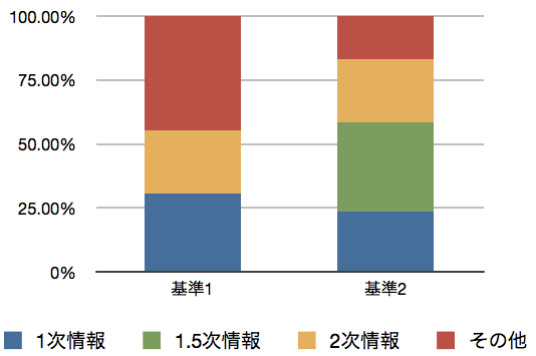

図 2: 震災ツイートログの分類

1.5 次情報を追加することで基準 1 よりもその他に分類されるツイートの件数が減少した．その他には，「そういえば卒業式どうなるのだろう。」といったひとりごとのようなものや，「@○○いってらっしゃーい！」という挨拶が存在した．

4. 考察

災害発生直後は，初動が重要であるため，正確な情報のみを迅速に提供する手段が求められる．したがって，1 次情報を最優先に判別する基準 1 が有効である．しかし，災害発生からある程度時間が経過すると物資配分や人力派遣など意志決定を必要とする状況に推移すると考えられ，この場合は基準 2 が有効となってくる．このように，目的やデータに応じて臨機応変に基準を使い分ける必要がある．これは情報トリアージそのものと言える．

認知バイアスの影響を確認するために，1 次情報の表層的な表現に注目する．総選挙ツイートログ (基準 1) の 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 (6.8%)，ニュートラルなツイートは 771 件 (77.1%)，ネガティブなツイートは 161 件 (16.1%) であった (図 3)．

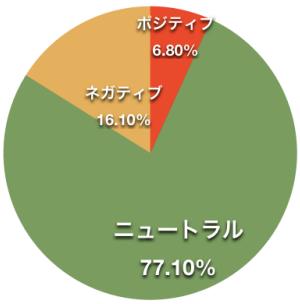

図 3: 総選挙ツイートログの極性分類

ポジティブなツイートには「○○党応援してます!!」等，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」等，特定の政党や候補者の悪いイメージを与えるもの，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等があった．

ニュートラルなツイートは中立的意見であり，有権者が投票する際の判断材料として妥当である．認知バイアスの影響を考えるとニュートラルなツイートのみが 1 次情報に出現するのが望ましい．そのため，複雑な意志決定プロセスを伴う場合には基準 1 の分類は望ましくない．それに対し，意見情報の分類定義が基準 2 は明確化されているため，1 次情報に認知バイアスの影響を与えるものが存在せず，複雑な意志決定プロセスを要する状況に利用するのに望ましい．

また，基準 2 による総選挙ツイートログの分類結果を見ると，1 次情報における意見や意思といったツイートが基準 1 よりも除去できていた．1.5 次情報を追加することで，1 次情報に存在する認知バイアスの影響を除去することができ，参考となる意見情報等を明確に判別して提示することが可能となった．震災ツイートログの分類結果も総選挙ツイートログと同様に 1 次情報から意見や意思といったツイートが除去できていた．そのため，今後は基準 2 の定義をベースに，緊急を要する決断に伴う場合は 1 次情報のみを提示するというような，

表 3: 分類基準 (基準 2)

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
意志表示		○		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

状況によって提示する情報の幅を変えることで対応できる基準を作るべきだと考えられる．

5. おわりに

意志決定・状況判断をするための情報ソースとしてツイッターを用いる状況を想定し，正確性が高く均一性のある情報を自動抽出する手法の提案を目指したツイート分類基準を定義した．また，複雑な意志決定プロセスに伴う基準では認知バイアスの影響を考慮し，新たに 1.5 次情報を定義した．

定義を基に総選挙ツイートログ，震災ツイートログを分類した結果，本研究の目的である情報トリアージの定義に則した分類を行うことができ，分類基準の有効性を確認した．さらに，対象とするデータによって必要な情報に違いが生じること，時間経過や目的に応じた基準の使い分けが必要であることが明らかとなった．

今後，情報トリアージとの関連を重視しながら，自動抽出手法について研究を進める予定である．

謝辞

本研究では，東日本大震災ビッグデータワークショップにおける提供データおよびタイムラインからの抽出データを利用しています．Twitter Japan 株式会社に感謝致します．

参考文献

[岩木祐輔 2009] 岩木祐輔. ”アダム ヤトフト, 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,”.” 電子情報通信学会・日本データベース学会・情報処理学会第 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), A6-6 (2009).

[風間一洋 2010] 風間一洋，今田美幸， and 柏木啓一郎：”Twitter 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).

[福島裕斗 2013] 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：“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”，情報処理北海道シンポジウム，講演論文集 p73-p78, 2013-10

[桑野孝光 2012] 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)

[梅島彩奈 2011] 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

[立入勝義 2011] 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)

[宮部真衣 2011] 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

[Catherine C. Marshall 1997] Catherine C. Marshall, Frank M. Shipman, III：”Spatial hypertext and the practice of information triage”，HYPERTEXT ’97 Proceedings of the eighth ACM conference on Hypertext p124-133

[Sofus A. Macskassy 2001] Sofus A. Macskassy, Haym Hirsh, Foster Provost, Ramesh Sankaranarayanan, Vasant Dhar：”Information Triage using Prospective Criteria”，Appears in User Modeling 2001 Workshop:Machine Learning, Information Retrieval and User Modeling

[Macskassy 2011] Macskassy, Sofus A. and Provost, Foster：”Intelligent Information Triage”，Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval，p318-326

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