E-014

Twitter を用いたニューストピックにおける少数意見の抽出
Extraction of Minority Opinion in News Topics using Twitter

熊田 研治 †　久保田 稔 †
Kenji KUMADA　Minoru KUBOTA

1. はじめに

近年，インターネット環境が身近になり，特にスマートフォンの普及，及びソーシャルメディアの登場により大量の情報が入手できる環境が普及してきた．しかし，大量情報の中では，多数意見やマスメディアの情報が多くを占め，別の視点による意見が見つけ難く，ユーザにとって必ずしも適切な情報が得られないことがある．本研究では，ニューストピックに関する情報発信における少数意見に注目する．ソーシャルメディアとして，ユーザが不特定多数に対し自由に自分の意見を投稿できる Twitter のツイートを対象とし，形態素解析と統計解析を用いて特定のニューストピックに関する少数意見の抽出を試みる．

2. 関連研究

Web 上のテキストデータから特徴語や傾向などを抽出及び分析する研究は多くある．ユーザが書いた映画レビューから TF-IDF を用いて映画推薦する研究[1]や，Twitter 上のツイート内容から TF-IDF を用いて特徴語を抽出し，ユーザの再発信行動を予測する研究[2]がある．また，N-gram を用いて特徴語を抽出し，ツイートの感情分析[3]したものがある．しかし，少数意見を対象としたものはない．本研究では，TF-IDF と N-gram を用いてツイートデータからツイートの傾向及び少数意見の抽出を行う．

3. 提案手法

一般に，大量のテキストデータを扱う場合，形態素解析を行って term（形態素の原形）を高頻度順に並べるとべき乗則になることがわかっている．

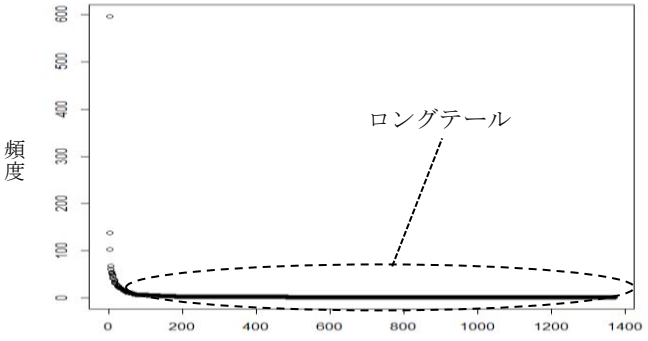

図 1　term の頻度分布

図 1 は，実験で用いたツイートデータの term を高頻度順に並べ替えた頻度分布である．図 1 から一部の term が出現頻度の上位の大半を占めているのがわかる．また，べき乗則を対数変換すると線形で表すことができジップの法則が示せるとの結果もある[3]．多数意見の場合，高頻度な term は数が限られており，これを手掛かりに傾向を読み解くことができる．しかし，少数意見の場合，ロングテール（低頻度な term）に含まれている可能性が高い．

そこで，少数意見を抽出するために，以下の手法（(I)～(IV)）を提案する．ここで少数意見とは，ニューストピックに対する個人的意見を述べているものと定義する．

(I) ツイートデータを形態素解析し term の頻度分布及び N-gram（bi-gram）解析の頻度分布から分類分け．
(II) 分類ごとの文長の分布を算出．
(III) 文長ごとの違いを比較し，ツイートの TF-IDF 値の合計値及び，N-gram の頻度の合計値を算出．
TF-IDF の式は，

$$tf(t,d)\times idf(t) \quad (1)$$

$$tf(t,d)=w_t^d \quad (2)$$

$$idf(t)=\log\frac{N}{df(t)}+1 \quad (3)$$

となる．w_t^d は，文書 d 内の term t の出現数であり，N は，全ツイート数，$df(t)$ は，term t が出現するツイート数である．

さらに，本研究では，term の頻度を $uf(t)=w_{i,j}$ とし，新たな重みを $ERf(t)$ としたとき，以下の式を定義する．

$$ERf(t)=\frac{1}{uf(t)}=\left(\frac{1}{w_{i,j}}\right) \quad (4)$$

$ERf(t)$ は，ツイート文に同じ文字が多数（「・・・選挙選挙選挙・・・」など）出てくると TF-IDF 値が高くなってしまうのを防ぐためである．$ERf(t)$ で求めた重みを（1）式にアダマール積とすると，

$$\{tf(t,d)\times idf(t)\}\circ ERf(t) \quad (5)$$

となる．（1）式を TF-IDF，（5）式を TF-IDF_{er} とする．

(IV) 上記の TF-IDF，TF-IDF_{er} の合計値から少数意見の抽出割合を算出．

4. 実験

解析対象のツイートデータは，2014 年 12 月 14 日（日）に行われた衆議院議員総選挙とし，時間帯は，開票直前の 20 時前とした．ツイート検索語は，「選挙」とし，ツイートデータ収集は，株式会社アシストが提供する Twitter Connector（2015 年 1 月 30 日をもって提供は終了している）を用いた．今回は，一般ツイートを対象とし，http，@，# を含むツイートは除去している．

形態素解析及び N-gram 解析の対象は，名詞・動詞・形容詞とした．また，顔文字や記号・特殊文字は，辞書の都合上名詞と判断される場合が発生してしまうため，半角空白に置換してある．N-gram は分類分けの精度が高い[3]と

† 千葉工業大学工学研究科
† Graduate School of Engineering, Chiba Institute of Technology

されていた bi-gram を用いる．分類分けは人手で行うためデータ数は 500 件とした．

5. 評価

形態素解析の高頻度 term 及び bi-gram の高頻度の代表的なものを表 1 に示す．

表 1　代表的な高頻度

	高頻度←				
term	する	テレビ	見る	番組	面白い
bi-gram	選挙-速報	池上-さん	テレビ-全部		

高頻度なものからツイートデータを抜き出して見ていくと，主観ではあるが以下のような 5 分類に分けることができる（G1～G5 はグループ名）．

G1：選挙関連の一般的ツイート（選挙行く，投票する，選挙見るなど）と選挙情報（244 件）．

G2：つまらない（録画・DVD 見る，選挙ばっかりなど）（117 件）．

G3：おもしろい（41 件）．

G4：少数意見（59 件）．

G5：不定（39 件）．

G4 には，「日本人って実は普段は他者を貶して・・・野党の皆様方は好んで自民の批判しかせず・・・」や「・・・個人的には、投票に行かないという選択肢もありだと思う・・・私が選挙行かないときは、国のやり方や流れを変える必要性を感じない時だわ」などがある．また，G5 には，「選挙権まであと 5 年」や「選挙のせいかツイッターおもい」などである．

ここで提案手法の前に，一般的分類法であるクラスタ分析を試みた．クラスタ分析には，階層型と非階層型があるが，データ数が多い場合，階層型は扱いづらいので非階層型である k-means 法を用いた．分類結果を表 2 に示す．

表 2　k-means 法による分類結果

グループ	k1	k2	k3	k4	k5
件数	2	361	25	1	111

表 2 からグループ k1 とグループ k4 の件数が極端に少ないのがわかる．ツイートの中身を見てみると「～・・選挙選挙選挙・・～」というツイートであるのが確認できた．これにより，特殊なツイートがあるとそれだけでグループを 1 つ使用してしまうという欠点がある．ツイートの場合，曖昧性やユニーク性，遊び言葉など様々な言葉が入ってくるため通常の分類では上手くいかないことがわかる．次に文長の長さを比較するためにグループごとに term 数の比較をする．結果を図 2 に示す．

図 2 の結果，少数意見である G4 は，他のグループよりもツイート文が長くなる傾向があることがわかる．この結果を踏まえ，TF-IDF と TF-IDF_{er} 式の値の合計と bi-gram の頻度を高い順に並べ替え，上位 5%～上位 50%で少数意見が何件含まれているかを解析する．結果を表 3 に示す．

表 3 の結果，TF-IDF と TF-IDF_{er} 式の値の合計の上位 5%で約 60%，上位 10%で約 50%が少数意見であることがわかった．bi-gram の頻度合計値の上位では，TF-IDF ほどではないが，上位 5%で約 54%，上位 10%で約 45%が少数意見である．TF-IDF，TF-IDF_{er} 式では，ほぼ同じ結果となったが，これは，5%，10%などの区間内で件数の変化が

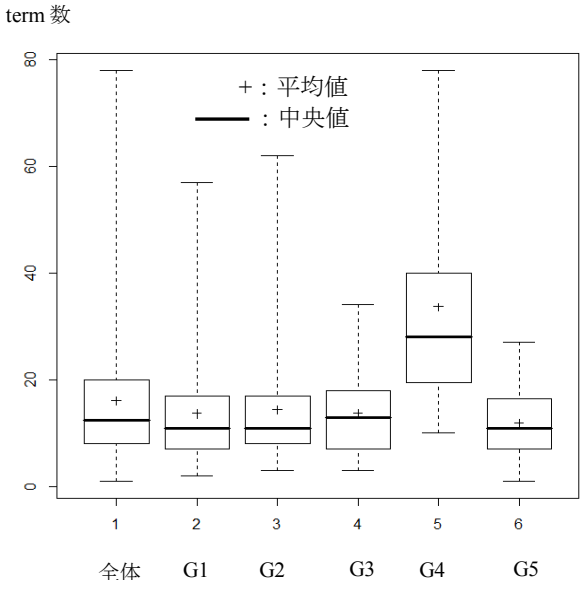

図 2　G1-G5 の分布

表 3　少数意見の比率

	TF-IDF		TF-IDF_{er}		bi-gram	
	件	比率	件	比率	件	比率
上位 5%	16	0.64	15	0.6	14	0.538
10%	25	0.5	27	0.5	23	0.451
20%	39	0.39	41	0.402	37	0.366
30%	48	0.32	48	0.32	47	0.281
40%	53	0.265	55	0.266	53	0.235
50%	56	0.224	57	0.223	57	0.217
全体	59	0.118	59	0.118	59	0.118

見られなかったことであり，TF-IDF_{er} 式を用いることで少数意見は区間内で順位を上げていたことを確認できた（一部逆行していた場合もある）．これにより，TF-IDF_{er} 式のほうが少数意見を抽出しやすいと推測する．

以上から，ツイートを分類するには N-gram を手掛かりに分類を行い，$ERf(t)$ を用いた TF-IDF_{er} 式を用いることで少数意見を抽出しやすくなると考える．

6. まとめ

少数意見の傾向（今回は文長の長さ）を手掛かりにし，また，同じ文字が多数出現する場合の対策として $ERf(t)$ を定義し，TF-IDF_{er} による少数意見の抽出を試みた．本提案手法では，TF-IDF_{er} で得られる値の上位区間に少数意見が含まれている可能性が高いと示しただけで，曖昧性も残った．今後は，少数意見の term の発生頻度から少数意見の抽出を行っていく．

参考文献

[1] 林 貴宏，尾内 理紀夫，“Web 上のレビューを利用した映画推薦システム”，人工知能学会論文誌，Vol.30，No.1，pp.102-111，2015.1.

[2] 阿部 秀尚，“Twitter 上での発話履歴の時系列パターンに基づく特定発話行動予測手法の検討”，情報処理学会研究報告知能システム(ICS)，2015-ICS-178(11)，pp.1-5，2015.2.

[3] Alexander Pak，Patrick Paroubek，“Twitter as a Corpus for Sentiment Analysis and Opinion Mining”，Proc. LREC 2010，pp.1320-1326, 2010.