3Z-04

ソーシャルメディアにおける少数意見の抽出

熊田　研治†　　久保田　稔†
千葉工業大学大学院工学研究科†

1. はじめに

近年，インターネット及びスマートフォンの普及，さらに SNS 利用者の増加により大量の情報が手に入る環境にある．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い．本研究では，ニューストピックに関する情報発信の少数意見に注目する．SNS としてユーザが不特定多数に自由に自分の意見を投稿できる Twitter を対象とし，形態素解析と統計解析を用いてニューストピックに関する少数意見の抽出及び分類を試みる．

2. 関連研究

Twitter 上での発言に関し，多数派認知とフォロワー間の研究[1]がある．フォロワー間での発言は，同質性があり多数派に傾向するとしているが，オリジナル tweet には同質性は確認できなかったとある．トピックの種類によってユーザの発言は変わることもあり，本研究で目標とする少数意見（少数派）の抽出は重要である．

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）の頻度が多い順に並べると，べき乗則になることがわかっている．

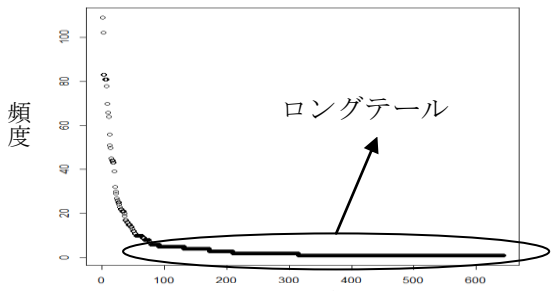

図 1 term の頻度分布

図 1 は，実験で用いた tweet データの term の頻度分布である．高頻度の term から多数意見の傾向を読み解くことはできるが，少数意見は，ロングテールに含まれている可能性が高い．また，低頻度の term には，記号や本来の意味とは全く関係ない term などのノイズも含まれる．

少数意見を抽出するために，以下の手法を提案する．

(1) tweet データを形態素解析し term の頻度分布及び動詞頻度を抽出．
(2) term の高頻度から Web 情報と多数意見を把握．高頻度順に並べニュース情報及び多数意見を決定．
(3) tweet データに動詞がない場合，Web ニュースのみと決定．動詞のない tweet データを抽出すると Web ニュースの見出しだけである可能性が高いことを確認．
(4) 上記以外の Web ニュースのみの tweet データを取り除き，また，リンク情報などのノイズを人手で削除し，クラスター分析で少数意見を分類．

4. 実験

Twitter API を用いて取得した tweet データを用いて，提案手法の有効性を確認する．ニューストピックは，政治，経済，スポーツを選定した．3 つのトピックのキーワードは，Yahoo!JAPAN のニュース見出しから抽出した．「アメリカ中間選挙のニュース」では，ニュース見出しが「オバマ氏敗北 政権運営厳しく」であり，キーワードの手掛かりとして「オバマ」と「政権」とした．以下のⅠ，Ⅱ，Ⅲは，実験対象のトピックに対応した実験番号で，→の後が抽出したキーワードである．

Ⅰ「アメリカ中間選挙のニュース」→オバマ&政権
Ⅱ「松坂大輔ソフトバンク入団のニュース」→松坂大輔&ソフトバンク
Ⅲ「吉野家の値上げのニュース」→吉野家&値上げ

上記 3 つの実験毎に tweet データを 100 件取得した．

5. 評価

今回の実験では，「オバマ政権押し」のよう

Extraction of Minority Opinions in Social Media
†Kenji KUMADA, †Minoru KUBOTA
†Graduate School of Engineering, Chiba Institute of Technology

な短文は少数意見に含めていない．オバマ&政権での高頻度 term の上位を表 1 に示す．

表 1　頻度表

rank	term	freq
1	民主	50
2	歴史	43
3	大敗	43
4	共和党	30
5	する	29
6	厳しい	23

表 1 の結果から，上位の term から tweet データを抽出すると，「失点続き」，「アメリカが腑抜け」などオバマ政権への苦境や批判の tweet が多く，Web 情報や多数意見であると判断できる．

動詞の頻度を表 2 に示す．横軸が頻度，縦軸が実験種別である．

表 2　動詞の頻度

個	0	1	2	3	4	5	6	7	8	9	10
Ⅰ	51	9	8	6	6	3	8	3	6	0	0
Ⅱ	63	0	23	9	3	0	1	1	0	0	0
Ⅲ	22	29	10	17	4	6	4	2	1	0	5

表 2 の結果，実験Ⅰから，動詞 0 個の tweet データ 51 件の内容は，「[国際] 米中間選挙，民主が歴史的大敗 オバマ政権、戦略漂流 （共同通信）」などである．Tweet データに動詞が含まれない場合，そのデータは Web ニュースであることが推定でき，少数意見抽出データから除くことができる．

少数意見を抽出及び分類するために，canberra 距離による ward 法でのクラスター分析した実験Ⅰの結果を図 2 に示す．クラスター間の tweet データの内容を分析すると，A，B，C，D，E，F の 6 つに分類できた．A には，「安倍首相も消費増税すると同じ轍を踏む」，B「オバマ政権にはあまり知日派はいない」，C「好景気の恩恵を受けない」，D「ろくでもない政権へ」，E「中国，ロシア、イスラーム国問題は，オバマ政権が腑抜けだから」，とオバマ政権への批判的なものと見ることができる．F については，「公務おつかれさま」，「ルビオ大統領の登場を待つ」，「イスラエルの思うつぼ」など 8 件あり，個人的意見でまとまっていた．内容は，単なるオバマ政権への批判ではなく，ユーザ個人のニューストピックに対する発言であり，少数意見と判断できる．

図 3 に示す実験Ⅱの分析結果では，A は「決まった」，「もう 34 歳だったのか」などの情報意見であり，B は「ホークス参戦前は西武か横浜かとおもったのになぁ」，「ローテーション確認して行かなければ」，「私にとって松坂大輔と中村紀洋はかつて最も好きだった選手」など 12 件の個人的意見があり実験Ⅰ同様に少数意見であると判断できる．

実験Ⅲでは，Web ニュースのみの情報が少なく個人的意見が多かったため，図（デンドログラム）は省略した．少数意見として「マックより高いのか」，「今までが異常な安さだった」，「国産メニュー作って」，「ラーメン一杯 800 円のほうが高い」など 31 件があった．

キーワード：オバマ&政権

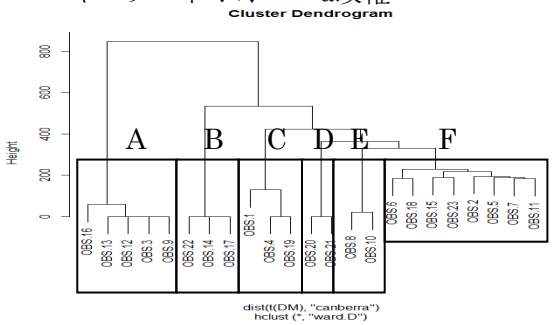

図 2　クラスター分析例（実験Ⅰ）

キーワード：松坂大輔&ソフトバンク

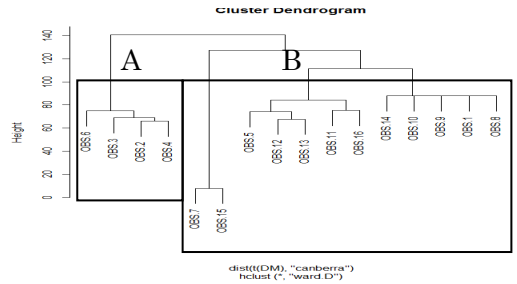

図 3　クラスター分析例（実験Ⅱ）

6. まとめ

本稿では，少数意見を抽出するため tweet データの形態素頻度と動詞の数に注目し，tweet データの絞り込みを試み，残った tweet データをクラスター分析で少数意見の分類を行った．今後は，大量データに適した少数意見の分類法と時系列データを用いた少数意見の発生頻度に関する研究を進めていく．

参考文献

[1] 小川　祐樹，山本　仁志，宮田　加久子，”Twitter における意見の多数派認知とパーソナルネットワークの同質性が発言に与える影響，” 人工知能学会論文誌，Vol.29, No.5, pp.483-492, 2014.8.