1F4-3

Web 上の政治的発言とその応用

Classification of the political opinion on the web and its application

東　宏一　　掛谷　英紀
qq274sw9k@yahoo.co.jp　　kake@iit.tsukuba.ac.jp
筑波大学

Abstract: In this paper the author tries classification of Japanese diet members by using linguistic features of tweets. Maximum Entropy Method is used for splitting the all tweets into political opinion and daily life. Self-Organizing Maps is used for the classification of the tweets within each category and a map with five main clusters is generated as a result.

1. はじめに

近年、国政選挙などにおいて各政党が『マニフェスト』を発表し、その内容を比較して有権者が投票したい政党、議員を選ぶというスタイルが一般化しつつある。これに伴い、マニフェストに対する有権者の評価に基づいて、投票すべき政党を推薦するシステムも提案されている[1]。これは投票支援システムを実現する試みであるが、課題も存在する。それは、マニフェストは個々の政党にとって基本的な戦略ではあるが、必ずしもその通りに実行される保証はないという点である。例えば、民主党のような多様なバックグラウンドを持つ議員が所属する政党においては、党内での政治志向の違いが大きく、全体としての一貫性がない。そのため、提案されたマニフェストが政党に所属する議員の総意であるとは言えない可能性も高い。

そうした問題点を解決して投票支援を行う上で、先行研究として、みんなの政治を使った自己組織化マップがある。議員をイデオロギーに応じて分けることにある程度成功しているが、あくまでも不特定多数のネットの書き込みをベースにしたものであるため、信頼性という面で十分とはいえない[2]。

そのため、本研究では次のような手法を提案する。まず、有権者が普段目にする、政治的な話題に明るい人々として著名な知識人が考えられる。これらの人々はメディアを通し、自らの政治志向に基づいた発言を行っているので、有権者にとって個々の知識人の考えは把握しやすい。そこで、知識人と議員との間の政治志向の類似度を調べ、それを可視化することができれば、選挙の際に有権者が自身の考えに近い候補者を選ぶことが容易になると考えられる。

このことを具体的なシステムとして実現する手法として、自己組織化マップに知識人と議員の発言を入力して分類を行うというものが考えられる。ただし、前提として議員の発言に対する分類の正当性が保証されている必要があるので、その検証も行っていく。また、発言を収集するための情報源としては、ツイッターに着目している[3]。

本研究では、投票支援システム構築の予備段階として、ツイッターのアカウントを持つ現職国会議員を対象として、その発言傾向の類似度を自己組織化マップにより可視化することを目指す。

また、国会議員のツイッター上での発言には、日常生活に関するものと、議員自身の政治的な信条に基づくもの（以下、政治的な話題と呼ぶ）とが混在しており、これらを分離して政治的な話題のみを抽出した上でマップを出力することも目指す。

最後に、出力したマップ上での分類の正当性を、マップ上の代表ベクトルの要素である、素性に対する分析により検証する。

2. システムの概要

本研究では、形態素解析ツールとして、ChaSen を用いる[4]。まず ChaSen を用い電子化されている複数の文書データを形態素解析して名詞のみを抽出し、得られた素性から学習データを作成する。

学習データを元に、機械学習のプログラムで文章の特徴を学習し、分類を行う。機械学習には自己組織化マップのアルゴリズムを用いる。今回、自己組織化マップの作成プログラムとしては、市販本に付属のプログラムを利用した[5]。学習データの生成から自己組織化マップを作成するまでのシステムの流れを図 1 に示す。

また、今回全てのツイートの中から政治的な話題のみを抽出するために、最大エントロピー法による機械学習を行った［6]。

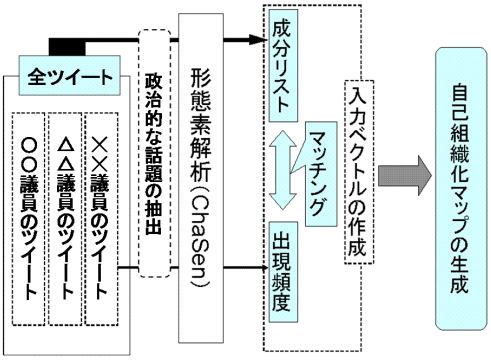

図 1　システムの概要図

3. 学習データの作成

3.1 国会議員のツイッターの収集

ツイッターは米国ツイッター社が提供する SNS サービスである[2]。日本でも多くの人々が利用しており、国会議員にも多くの利用者が見られる。今回このツイッターを学習データとして利用したのは、ブログなどに比べ簡便なため、各人のその時々の率直な心象・意見などが表現されていることが期待されるためである。

本研究では、2010 年 12 月時点でツイッターのアカウントを取得している国会議員 61 名を対象とした。また、今回使用したツイートは、発言が 40 字を超えるもののみに限っている。

3.2 学習データの作成

自己組織化マップを作成するためには、収集した学習用データを入力ベクトルの形に加工する必要がある。本研究では、入力ベクトルの要素として素性の出現頻度を用いている。

まず、全議員のツイートに含まれる素性を出現件数の降順に並べ、上から 2000 個を抽出する。こうして作成されたリストを参照リストと呼ぶ。次に、参照リスト内の素性 α が、各議員のツイート内でどのくらい出現しているかという頻度 $fre(\alpha)$ を算出する。ここで、この算出式は、

$$fre(\alpha) = \frac{n_\alpha}{\sum_i n_i} \qquad (1)$$

のように表される。式(1)において、 i は各議員ツイート内に含まれる素性であり、 n_i はその出現件数である。また、素性 α がその議員のツイートに含まれていない場合、頻度は 0 となる。

4. 実験

4.1 政治的な話題の抽出

政治的な話題を抽出するために、以下の手法を用いた。

① ランダムに抽出した 800 件のツイートに対し、手動で政治/日常のいずれかに分類する。

② ①で作成した教師信号に基づいて、全てのツイートを最大エントロピー法によって政治的な話題と日常生活に関する話題の 2 つに分類する。その際に政治的な話題として類されたツイートを用いて、自己組織化マップを出力する。

作成した教師信号に対して 10 分割クロスバリデーションによるテストを行ったところ、正解率は 78.42%であった。

4.2 政治・日常を分離せずに実験

作成した学習データについて、入力ベクトルが 2000 次元のときの自己組織化マップを出力した。所属政党は各 ID の最初の英字で示している。英字は政党名を表し、M は民主党、K は公明党、J は自民党、S は社民党、Mi はみんなの党、SN は新党日本、SK は新党改革である。また、マップ上のグレースケールはノード同士の距離を表しており、黒に近いほどノード間の隔たりが大きいことを示している。

政治的な話題と日常生活の話題とを区別せず、マップを出力した場合の結果を図 2 に示す。

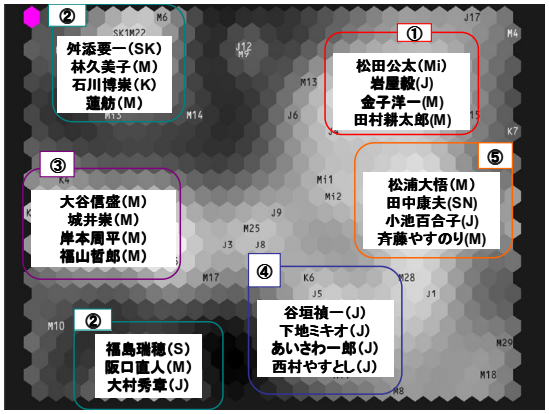

図 2　出力マップと 5 つのクラスタ

クラスタ間での発言傾向の違いを知るため、それぞれの中心となるノードにおいて、ベクトルが持つ要素を分析した。その結果得られた素性の例を表 1 に示す。素性の抽出は以下の手順で行った。

① まず、各代表ベクトルにおいて素性の出現頻度(tf)が平均を上回っており、かつその差分が大きいものを抽出する。ここでいう出現頻度の平均とは、全ての議員のツイートをまとめた上で式(1)の計算を行った際の、各素性の出現頻度である。

② 次に、クラスタごとに特徴的な素性を抽出するため、①で得られた各クラスタの素性に idf 法を用いてソートを行う。

この結果として得られた素性を降順に並べたのが表 1 である。これを見ると、「行革」「財特法」など政治に関わる素性も見られるが、食事や議員が出席するイベントなど日常生活に関わる話題も多い。

表 1　クラスタごとに特徴的な素性の例

①	②	③	④	⑤
静寂	レフト	京阪	急激	量的
なまず	ゲスト	ごまめ	尊敬	トビ
いつ	山菜	今日	イチ	朝市
すき間	予算	活動	会	志向
考え	祭り	寿司	独裁	充満
愉快	恵比寿	会	今日	妖星
記事	衣類	サナダムシ	鯛	老朽
出身	季刊	社会	料理	負担
偽者	税務署	アジェンダ	新幹線	スポーツ
感性	ピンク	主筆	喫茶店	ニックネーム
モーニングコート	社会	帰宅	用船	未払い
熱帯魚	工科	預託	肝臓	斎
お好み焼き	鹿	泣き	略	通商
アイデア	河川	儀	ボーイスカウト	けんぽ
エッセンス	行動	持ち場	ストリート	おと
新幹線	発言	日	札所	ハイボール
付与	ビジネスマン	ピンク	本件	プレゼント
拒否	ご免	衣類	行革	対等
減価	活動	ノー	風力	財特法

4.3　政治的な話題のみでの実験結果

次に、最大エントロピー法によって政治的な話題のみを抽出した上で実験を行った。その結果を図 3 に示す。

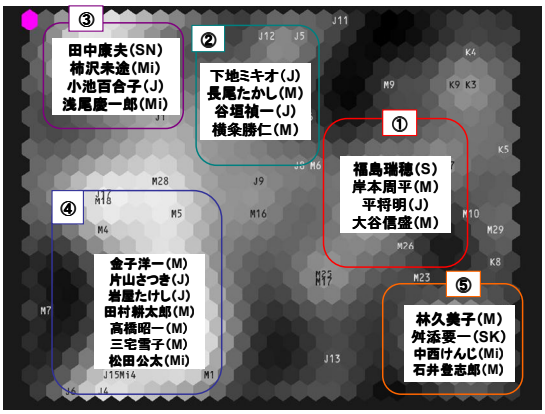

図 3　出力されたマップと 5 つのクラスタ
（政治的な話題を抽出した上で出力）

また、前節と同様に各クラスタの代表ベクトル間で素性を比較した。この結果を表 2 に示す。

これを見ると、「ブリーフ（ブリーフィング）」、「予算」、「ドラッグ」、「治水」など、前節に比べて全体的に政治に関わるキーワードは多く見られ

る。また、収集したツイートの中に尖閣諸島問題に関するものが含まれていたため、「領海」、「海域」などの言葉も見られる。

表 2　特徴的な素性の例（政治的な話題のみ）

①	②	③	④	⑤
愚か	メトロ	領海	ブリーフ	空襲
意志	国籍	紛れ	浜	セット
朝	立論	女性	竜	姿勢
料率	駆け込み	ポケットマネー	上京	動議
健全	金山	外側	直滑降	だらし
靴下	予算	海域	公共	直滑降
財界	理論	予備	治水	一読
懇談	パラリンピック	鶴	分断	異臭
地獄	早晩	次官	明日	浅瀬
顛末	肢体	ミイラ	バラマキ	引き下げ
含み	私見	乗員	選挙	県連
イベント	多く	一元化	クラブハウス	リスト
夕飯	前	一蹴	一人	意志
人当り	スムーズ	待ち人	保証	参議
多く	苛酷	訪韓	ネガティブ	全廃
感慨	治水	造形	催眠	死
崩落	汚染	超	私見	愛
ドラッグ	上京	危機	ウコン	強行
層	楽天	完全	入り	加害

5. おわりに

本研究では投票支援システムを作成するため、その前段階として、自己組織化マップを用いて国会議員をツイッター上での発言傾向によって分類した。政治的な話題を抽出してマップを出力した場合とそうでない場合を比較してみると、政治的な話題を抽出しない場合、全体的に議員自身の日常生活に関わる素性が多く見られた。

これとは逆に、政治的な話題を抽出した場合、議員自身の日々の政治活動に関わる素性や、政治的な課題に関わる素性が多く見られた。また、政治的な話題の中に、政治問題となった尖閣諸島問題に関する発言なども見られた。

ただ、議員がツイッター上で PR しているブログのタイトルなどが、出現頻度が高い素性として含まれてしまうケースもあり、これらを如何にして排除していくかが今後の課題となる。また、今回政治的な話題を抽出する際に用いた最大エントロピー法の正解率は 8 割弱であり、政治的な話題のみを抽出するという点ではまだ課題が残る。

また、分類された議員同士の政治的な立場を踏まえた上での評価なども、今後行っていく必要がある。

参考文献

[1] 静岡大学情報学部　佐藤哲也研究室 http://tai.ia.inf.shizuoka.ac.jp/

[2] 東宏一,橋本悠,掛谷英紀(2011), Web 上の言語資源に基づく自己組織化マップの作成,言語処理学会第 17 回年次大会

[3] twitter.jp http://twitter.com/

[4] 奈良先端科学技術大学院大学 松本研究室 ChaSen http://cl.aist-nara.ac.jp/

[5] 自己組織化マップとそのツール, シュプリンガージャパン, 大北正昭ら編

[6] 内元清貴, 村田真樹, 関根聡, 居佐原均(1999): 日本語係り受け解析に用いる ME モデルと解析精度, 言語処理学会第 5 回年次大会併設ワークショップ論文集.