Ｗｅｂ上の言語資源に基づく国会議員の分類

東　宏一　　橋本　悠　　掛谷　英紀
qq274sw9k@yahoo.co.jp　　s0920896@u.tskukuba.ac.jp　　kake@iit.tsukuba.ac.jp
筑波大学

概要：本研究では、Web上で公開されている政治に関するレビュー記事とツイッターに基づき、自己組織化マップによる国会議員の類似度マップを作成する。レビュー記事には、Yahoo!JAPANが運営するyahoo!みんなの政治のコンテンツ「みんなの評価」を採用し、ツイッターとしては同サービスのアカウントを持つ現職国会議員のツイートを収集した。この2タイプの文章データに対して、それぞれ自己組織化マップによる分類を行い、各クラスタにおける代表ベクトルの素性を比較した。その結果、各クラスタに所属する国会議員の政治志向を象徴すると思われるようなキーワードを見出すことができた。

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ために、次のような手法を提案する。まず、国会議員のレビュー記事としてYahoo!JAPANが運営する「みんなの政治」に着目する[2]。これは、現職の国会議員に対して有権者がレビューを投稿しているものであり、これを議員ごとに収集して分類することで、有権者が選挙の際に候補者を選ぶ参考になると考えられる。これをシステムとして実現するために、本研究では自己組織化マップに議員のレビューを入力して分類を行う。

また、議員レビューを用いる場合の問題点として、有名な議員にレビューが集中し、あまり知られていない若手議員に関する情報を得にくいというものがある。こうした議員に関する情報も含めた実用的な投票支援システムを構築するために、ツイッターに着目した[3]。議員レビューと同様に、現職国会議員のツイッターを収集し、同様に自己組織化マップによる分類を行う。

最後に、各出力マップの代表ベクトルに含まれる素性を調べることで分類の正当性を評価する。

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]。また、品詞は名詞のみを用いた。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]。システムの流れを図1に示す。

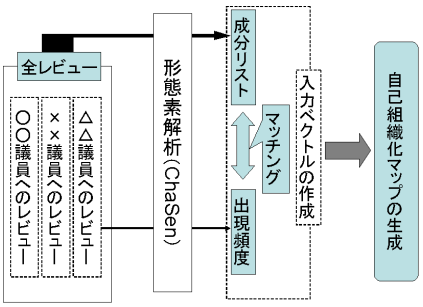

図1　システムの概要図

3. 学習データの作成

3.1 議員レビューの収集

今回、学習データの元データとして、2010 年 9 月に収集した「みんなの政治」内の自民党議員・民主党議員の否定的なレビュー記事(評価点 1、2 点)を用いる。否定的なレビュー記事に限定したのは、議員個人単位で素性の出現傾向を調べるためには肯定的なレビュー記事の絶対数が不足しているためである。素性の出現確率を求める際にレビュー記事数が少ない状態では単語の出現傾向を正しく評価することができないので、今回はその中から否定的なレビュー記事が 50 件以上投稿されている議員、計 83 名 17,020 件のレビュー記事を対象として実験を行う。

3.2 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。日本人の利用者も増加しており、国会議員にも若手を中心に多くの利用者が見られる。今回このツイッターも学習データとして用いたのは、議員レビューとは違い、若手議員の情報を豊富に集めることが可能だったためである。

本研究では、2010 年 6 月時点でツイッターのアカウントを取得しており、かつ 400 件以上のツイートを行っている国会議員 43 名についてツイートの収集を行った。また、議員個人の考えに基づく発言を収集することを目的としたため、ツイート内で“RT”（他人のツイートを引用して発言すること）以下の部分は削除した。

3.3 学習データの生成

学習に使用したレビュー記事 17020 件を形態素解析し名詞を抽出した後、それを基に各素性の全レビュー記事での出現確率を調べる。そして確率値の上位 1000 件にあたる素性を入力ベクトルデータの成分として採用する。つまり、生成するベクトルデータは 1000 次元となる。その各ベクトル成分に各議員の上記 1000 素性の出現確率を対応づけ、全議員のベクトルデータを生成する。ここで、各議員での素性の出現確率は、ある議員に投稿されたレビュー記事数を n 、 n 件中にある素性が現れた回数を m とすると、 m/n で表される。

また、ツイッターに関しても同様に学習データを生成したが、この場合は上位 5000 素性を用いた。

4. 実験

4.1 議員レビューでの実験結果

議員レビューについての分類結果を図 2 に示す。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、黒に近いほど隔たりが大きい。出力されたマップを見ると、大きく 9 つのクラスタに分類できることがわかる。図 2 に出力されたマップと各クラスタを示し、表 1 に各クラスタに含まれる議員名を示す。

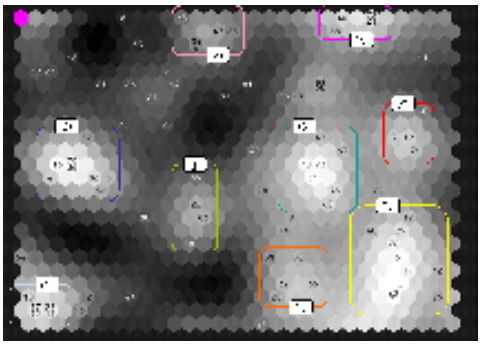

図 2　議員レビュー分類マップと 9 つのクラスタ

表 1　各クラスタに含まれる議員名

クラスタ	議員名(政党)　※自民党：自　民主党:民
①	田村幸太郎(自),横峰良郎(民),姫井由美子(民),三宅雪子(民),義家弘介(自)
②	森英介(自),小宮山洋子(民),生方幸夫(民),渡辺恒三(民),野田佳彦(民),仙石由人(民),前原誠司(民),枝野幸男(民)
③	佐藤正久(自),北沢俊美(民),赤松広隆(民),浜田靖一(自),石破茂(自)
④	管直人(民),藤井裕久(民),峰岸直樹(民),長妻昭(民)
⑤	小渕優子(自),塩崎恭久(自),山本拓(自),野田聖子(自)
⑥	二階俊博(自),古賀誠(自),平井たくや(自),山本有二(自),息吹文明(自),金子一義(自),江藤拓(自)
⑦	中川秀直(自),麻生太郎(自),森善朗(自),安部晋三(自),福田康夫(自)
⑧	大島理森(自),菅義偉(自),細田博之(自),村田吉隆(自)
⑨	大村秀章(自),松浪健太(自),世耕弘成(自),河野太郎(自),石原伸晃(自),後藤田正純(自),小泉新次郎(自),丸川珠代(自),平沢勝栄(自),山本一太(自),武部勤(自)

表２　図２の各クラスタでの頻出語句 20 件

①	②	③	④	⑤	⑥	⑦	⑧	⑨
当選	起訴	自衛隊	健康	少子化	宮崎	少子化	選対	出演
6	検察	防衛	手当	世襲	国交	宮崎	委員	チルドレン
番組	内部	軍	財務	民営	整備	惨敗	階	公認
学校	一致	基地	仕分け	厚生	族	健康	総務	TV
教育	仕分け	官	手当て	事故	国土	派閥	宮崎	番組
立候補	党内	処分	妻	落選	交通	総裁	疑惑	世襲
公認	捜査	普天間	税収	女性	道路	再生	長	息子
現場	日教組	戦争	子ども	歳費	産業	増税	補正	当選
離党	離党	給油	技術	郵政	東	国交	西松	若手
比例	逮捕	現場	無料	庁	建設	族	東	テレビ
被害	執行	宮崎	支給	公認	工事	大敗	報告	落選
出演	党首	米国	事業	出演	業者	整備	幹事	厚生
歳費	意図	他国	収入	担当	階	国土	下野	歳費
辞職	秘書	米	厚生	当選	知事	交通	世襲	離党
出馬	正義	沖縄	海外	如何	総務	改造	知事	立候補
先生	事件	平和	拡大	官房	県	会計	事務所	出馬
事情	組織	安全	高速	総務	高速	後期	審議	世
有権者	西松	派遣	削減	区	地元	構造	記載	次回
再生	交代	危険	発行	TV	公共	使い	制限	中身
所属	部	発生	産業	信念	業	高齢	規正	拝見

次に、各クラスタの代表ベクトルに含まれる素性の一部を表２に示す。表２の素性を見てみると、クラスタ③では議員の政治志向でなく、議員がこれまで経験してきた役職などに関連する言葉が特徴的な素性として挙がっていることがわかる。しかし、他のクラスターでは、たとえばクラスタ②は汚職に関係して批判されている議員群、クラスタ⑥は族議員として批判されている議員群、クラスタ⑨はテレビ等でのパフォーマンスが目立つとして批判されている議員群であるといったように、それぞれの議員をその政治志向によって分類できていることが見て取れる。

４．２　ツイッターでの実験結果

次に、ツイッターを用いた自己組織化マップの作成を試みた。その結果得られた自己組織化マップを図３に示す。各議員の所属政党は各 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 はみんなの党である。また、議員レビューと同様に、ツイッター分類マップを 5 つのクラスタに目視で括り、それぞれのクラスタに属する議員を表３にリスト化している。

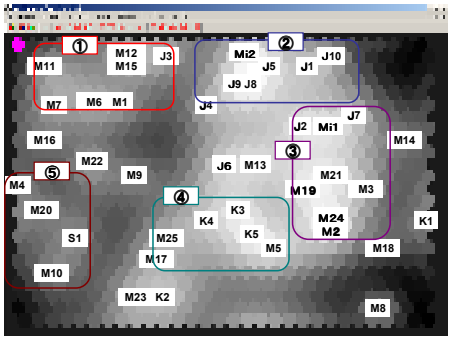

図３　ツイッター分類マップと５つのクラスタ

表３　各クラスタに含まれる議員名

クラスタ	議員名(政党)　※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
①	三宅雪子(M)、初鹿明博(M)、三村和也(M)、平将明(J)
②	逢沢一郎(J)、山本一太(J)、世耕弘成(J)、柿沢未途(Mi)
③	小池百合子(J)、岩屋たけし(J)、浅尾慶一郎(Mi)、金子洋一(M)、松浦大悟(M)
④	荒木きよぎろ(K)、谷合正明(K)、西田まこと(K)、斎藤やすのり(M)
⑤	逢坂誠二(M)、連舫(M)、山井和則(M)、福島瑞穂(S

さらに、議員レビューと同様に各クラスタ内の代表ベクトルの素性を表４に示す。表４の素性を見ると、民主メインのクラスタ①には特徴的な素性は特に見られないが、政権運営に関わる言葉が多い。自民メインのクラスタ②では、自民党に特徴的な『総裁』、他にも『基地』などの素性が見られる。政党混合のクラスタ③では、『デフレ』、『景気』など経済に関わる素性が見られる。公明メインのクラスタ④では、『青年』、『企業』など政党の特徴を反映すると思われる言葉が見られた。最後に、社民党を含む民主党議員のクラスタ⑤では、『子ども』、『女性』、『仕分け』など特徴的な素性が見られる。このことから、ツイッターを用いた自己組織化マップでも、議員の政治志向をある程度特徴づける分類ができていると考えれる。

表４　図３の各クラスタに特徴的な素性 20 件

①	②	③	④	⑤
会	選挙	の	街頭	今日
今日	党	こと	演説	今
会議	長	デフレ	挨拶	仕分け
委員	議員	ん	国政	者
議員	委員	的	報告	会議
今	会	もの	市内	子ども
政策	案	党	いま	事業
朝	問題	危機	方	息子
明日	質問	いま	企業	明日
厚生	の	朝立ち	皆さま	女性
朝立ち	予算	方	懇談	話
国会	総理	脱却	雨	集会
労働	国民	大変	会合	支援
地元	基地	出演	中	たち
会館	委	景気	号	皆様
意見	最新	候補	大変	朝
終了	本部	人	声	意見
ミーティング	候補	記念	青年	法人
駅	総裁	さん	きょう	会館
タウン	幹事	あと	皆さん	いつ

５．おわりに

本研究では、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、Yahoo!JAPAN が運営する「みんなの評価」の否定的なレビュー記事とツイッターのそれぞれに基づいて自己組織化マップによる国会議員の分類を行った。その結果、議員レビューに基づいて出力されたマップ上では、自民党の議員と民主党の議員が分離される形でそれぞれ集まっており、民主党・自民党という大きな括りでのイデオロギーによって分類できることがわかった。また自民党議員に比べ民主党議員はマップ上にバラついて配置されており、民主党が旧社会党の議員から旧自民党系の議員までを抱えており、党としてイデオロギー的統一感が弱いことがマップ上に表現されているのではないかと考えられる。また、加藤紘一氏が民主党議員側に入っているなど、思想を反映していると思われる部分も多く見られる。また、ツイッターに基づく分類結果では、自民党や公明党議員と同じクラスタに含まれる民主党議員も多く存在し、これも政党に所属する議員全体での一体感が乏しいことを裏付けるものではないかと考えられる。

しかし、ツイッターの場合には、議員個人の日常生活に関する話題と政治的な主張が混在しており、日常生活に関する素性をいかに排除していくのかが今後の課題となる。この問題を解決していくことで、より精度の高い自己組織化マップが形成されることが期待される。この解決策としては、日常生活に関する話題と政治的な主張をあらかじめ手動で分類し、それらを教師信号として全体をこの２つに分類するということが考えられる。

また、政治的な話題に詳しい方などに依頼して出力されたマップにおける分類の正当性評価を行っていくことも必要となる。

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] Yahoo!みんなの政治 http://seiji.yahoo.co.jp

[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