人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-10

地域特性を表すツイートの探索的閲覧支援システムの開発

Developing an Exploratory Tweet-Browsing System to Analyze Local Reputation Trends in Geographical Regions

森田 洋平[1]　白松 俊[1]　岩田 彰[1]

Yohei Morita [1], Shun Shiramatsu [1], and Akira Iwata [1]

[1]名古屋工業大学 大学院工学研究科
[1] Graduate School of Engineering, Nagoya Institute of Technology

Abstract: In this paper, we aim to develop an exploratory tweet-browsing system to analyze local reputation trends in geographical regions. We implemented functions for mapping the local reputation information, for extracting local feature words, and for exploratory browsing of local tweets. We conducted an experiment to evaluate whether users can extract local reputation trends about the snap election in 2014 by using our system. The experimental result indicated that users can extract local reputation trends about political topics in a particular prefecture by using our system. After this experiment, we also conducted a questionnaire survey about the usability of our system.

1. はじめに

近年，マイクロブログなどのソーシャルメディアが急速に普及してきており，個人が自分の意見を発信する機会が増えてきている．そのため，他の多くの人の意見や評判情報を得たいと思った時に手軽に収集可能となってきている．マイクロブログ上の意見や評判情報は投稿量が多く，様々な年齢，性別，地域の人々が投稿するため，アンケートやレビューでは得ることのできない情報を得られる可能性がある．本研究では特に，マイクロブログから地域に特有の意見や評判情報を得られる可能性に着目する．

毎日新聞では，立命館大学との共同研究プロジェクトで政党・政治家や有権者の Twitter 上のつぶやきやリツイート数，つぶやかれた単語などを分析するとともに，従来型の世論調査による結果内容との比較し，ネットでの呼びかけによる影響やユーザの関心などを調査している[1]．このように，新聞や雑誌においては世論調査，企業においてはマーケティング等にも利用されるなど，一般市民の日常の意見を抽出するための対象として非常に関心が持たれている．しかし，マイクロブログの利用方法が多様化していく中で，情報を収集・分析・可視化まで行えるアプリケーションというのは未だ少数である．また，世論調査やマーケティングなどにおいては，地域によって人々の意見が異なるため地域ごとに分析を行うことで新たな発見があると考えられるが，我々の知る限り，そのような研究は多くない．

そこで本研究では，日本国内での利用者も多く，大量の情報発信が行われているマイクロブログの一つである Twitter を対象に，世論調査や，マーケティング等に利用できるようにツイートを収集し，地域ごとに分析を行い，可視化するアプリケーションの開発を目指す．

2. 関連研究

2.1. 評判を抽出するための可視化の研究

Twitter を用いて意見・評判情報を可視化するアプリケーションとして，ヤフー株式会社が提供する，Twitter 上の投稿を検索できる「Yahoo!リアルタイム検索」において公開されている，つぶやき感情分析[2]がある．つぶやき感情分析は、検索したキーワードについてユーザがどのような感情を持っているかを、「ポジティブ」「ネガティブ」の割合でグラフ表示する機能である．

図 1 に示すように，ユーザが検索ボックスに語句を入力し検索すると，Twitter からその語句を含むツイートを収集し，画面左にストリーミング形式で表示する．画面右側には上から順に，ツイート数の時系列推移，ポジネガの割合，ポジネガの割合の時系列推移，トレンド語句が表示される．

また，膨大なツイートの中から評判情報を効率的

に取得するための可視化の研究として，村上ら[3]は任意の検索語句に対する意見がポジティブなものか，ネガティブなものかの評価をポジティブ度・ネガティブ度のキャラクターを用いた可視化と，評価極性ごとのツイート集合の表示，関連ワードの表示をするアプリケーションの開発を行っている．また，糸川ら[4]は，Twitter 上である話題に関する発言を分析，評価するための探索的ツイート閲覧システムについて提案している．

図 1　Yahoo!リアルタイム検索結果

2.2. 関連研究における課題・開発目的

前節で述べた関連研究のアプリケーションでは，政党に対する評判などの地域によって評判傾向が異なる題材に対しては人口が多く，投稿数の多い地域の評判傾向が全体の評判傾向として強く現れてしまい投稿量の少ない地域の特徴的な意見が埋もれてしまう危険性がある．

そこで本研究では，ある対象に対する評判を地域ごとに分析し，地域の特徴的な意見を抽出できるシステムの開発を目指す．

3. システム構成とインタフェース

本研究で開発したシステムの全体の流れを図 2 に示す．本システムでは，Twitter から TwitterAPI を用いてツイートを収集し，ツイートに対して前処理を行い，データベースに保管する機能を持つデータベースサーバと，クライアントから受けとったクエリーを含むツイートをデータベースサーバから受け取り，ブラウザに表示する情報の計算を行う，アプリケーションサーバで構成されている．

3.1. ツイートの前処理

3.1.1.　位置情報の取得

地域別で分析を行うため，ツイートの位置情報の取得が必要である．Twitter では，ユーザアカウントの設定からツイートに位置情報を付加することができる．しかし，筆者が収集したツイートに対して調査したところ，位置情報が付加されたツイートの数は 1%にも満たなかった．分析に用いるツイートのデータ数が少ないと意見に偏りが出る可能性や，有用な情報を得られない可能性がある．そこで本研究では，「ツイートの文章」・「ユーザプロフィール用の居住地」・「ユーザプロフィール」のテキストから位置情報の取得を試みた．

テキストからの位置情報の取得方法として，GeoNLP プロジェクトが開発している，ジオタギングツール GeoNLP[4]を使用した．GeoNLP は，自然言語テキスト内に含まれる地名や住所，施設名などにタグ付けを行い，そのタグに緯度・経度の情報を埋め込むツールである．

本研究で GeoNLP を用いて地名を解析する範囲は「ツイートの文章」と，ツイートした位置が，ユーザの居住地と近い可能性が高いことから「ユーザプロフィールの居住地」，「ユーザプロフィール」の 3 箇所とした．位置情報の優先度は，「ツイートに付加されている位置情報」＞「ツイートの文章」＞「ユ

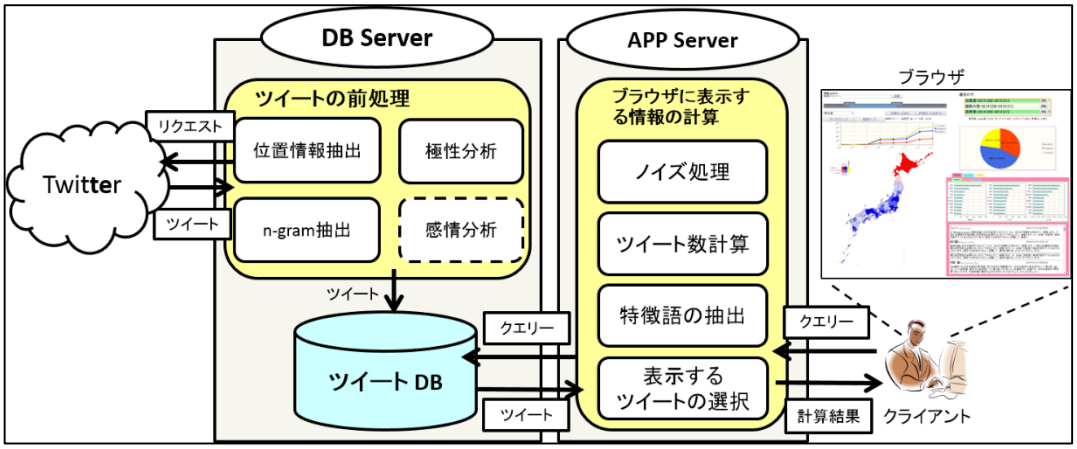

図 2：開発システム概要図

ーザプロフィールの居住地」＞「ユーザプロフィール」とした．また，地名語が複数あった場合は，得られた位置情報全てで，つぶやかれたツイートとして分析を行った．

本研究で収集したツイートに対して GeoNLP で位置情報の取得を試みたところ，全体の 30%に位置情報が付加された．

3.1.2. 評価極性分析

ある対象に関する評価を含むツイートを効率よく抽出するため，ツイートを評価極性で分類をする評価極性分析を行った．評価極性とは，その評価情報が「良い」や「好き」に代表されるポジティブな意味を持つのか，「悪い」や「嫌い」に代表されるネガティブな意味を持つのかを表す．本研究ではツイートをポジティブ・ネガティブ・評価なしの 3 種類に分類を行った．

評価極性の判定には，株式会社エクシングの提供する言語解析 WebAPI[6]を使用する．言語解析 WebAPI では「文単位」，「単語単位」のポジネガを判定することができる．

本研究では，ツイートを文で区切り，1 文単位で API の入力として，以下の条件で「ポジティブ」，「ネガティブ」，「評価なし」の 3 つの評価値にツイートを分類した．

1. ポジティブ・ネガティブに分類された文がない場合，ツイートの評価極性を「評価なし」と分類する
2. ポジティブ・ネガティブに分類された文がある場合，文数が多い方をツイートの評価極性とする
3. ポジティブ・ネガティブに分類された文の数が同数の場合，経験則により，より後半に出現する文の評価値を優先し，その評価値をツイートの評価極性とする

筆者が「ポジティブ」，「ネガティブ」，「評価なし」に分類した各 30 件，合計 90 件のツイートの文章に対し，言語解析 WebAPI を用いて分類した結果,正答率はポジティブが 76.7%，ネガティブが 80.0%，評価なしが 63.3%となった．

3.1.3. n-gram の抽出

n-gram とは，隣り合って出現した n 形態素を単語の単位とすることである．特に n=1 である n-gram をユニグラム，n=2 である n-gram をバイグラム，n=3 である n-gram をトライグラムと呼ぶ．本研究では，ツイートの文章がそれほど長くないことや，計算量を考慮してトライグラムまでを収集の範囲とした．n-gram の抽出のための形態素解析に MeCab[7]を用いた．

文章中に含まれる全ての n-gram を取得した場合，おそらく，その文章がどのような話題について書かれているか（政治に関するものか，スポーツに関するものか，など）を考える場合に有益とは思えない「は」，「が」，「です」といった助詞や助動詞，「ある」，「いる」，「あれ」，「それ」といった動詞や名詞でも，どのような話題の文章にも登場するような単語が特徴語として提示されると考えられる．このように，話題の種類と関連を持たないと考えられる単語のことをストップワードと呼ぶ．

本研究では，ユニグラムの場合はストップワードとなる語を全て除去し，バイグラム，トライグラムの場合はストップワードが語頭，または語尾に来る場合に除去した（一部例外を除く）．

また， 全ての n-gram に共通で接尾辞は一つ前の形態素と繋げて一つの形態素とした．接尾辞とは，「～さん」，「～的」といった単独では用いられず，常に他の語の下について，その語とともに 1 語を形成するもののことである．また，語尾の品詞が動詞であった場合は，原型に直して取得を行った．

3.2. インタフェース

ブラウザに表示されるインタフェースは図 3 である．インタフェースは，目的となる地域の特徴的な意見を含むツイートを探索的に閲覧できる用にするため以下の様な構成になっている．

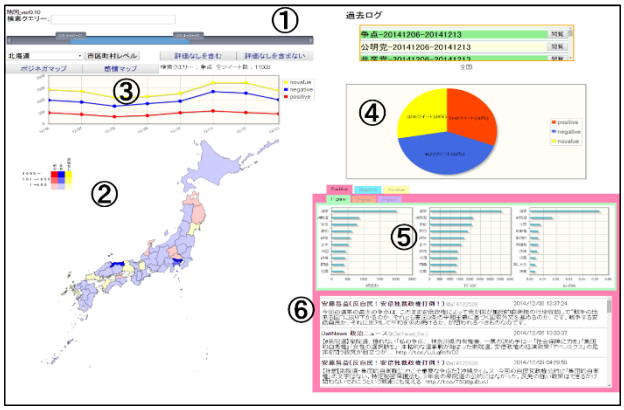

図 3:インタフェース全体図

① クエリー入力部

ユーザが分析したい情報のクエリーを入力する部分である．分析したいツイートの日付範囲を下部のタイムスライダーで入力することができる．一度分析したクエリーと日付範囲に対しては過去ログに追加され，分析を再度行うことなく閲覧することができる．過去ログには「クエリー」－「開始日」－「終了日」の順で表示されている．

② 評価極性マップ

評価極性ごとで一番割合の高いものを色で，ツイート数の多さを色の濃度で表現し，日本地図に可視化することができる．評価極性の色は，ポジティブが「赤」，ネガティブが「青」，評価なしが「黄」となっている. ツイート数は色の濃度で表現している．各都道府県をクリックすることで，クリックした都道府県のみのツイートを分析した結果（ツイート数の時系列グラフ，評価極性割合グラフ，特徴語グラフ, タイムラインビュー）を閲覧することができる．

また，都道府県選択後に「市区町村レベル」ボタンを押すことで市区町村レベルでの分析結果も閲覧することができる．

③ ツイート数の時系列グラフ

日付ごとのツイート数を，積み上げ折れ線グラフで閲覧することができる．横軸は日付，縦軸はツイート数，色は評価極性マップと同様である．また，グラフ上のマーカーをクリックすることで，クリックしたマーカーの日付のみのツイートを分析した結果（評価極性割合グラフ，特徴語グラフ，タイムラインビュー）を閲覧することができる．

④ 評価極性割合グラフ

現在分析しているツイート集合の評価極性ごとのツイート数と割合を閲覧する事ができる．色は，評価極性マップと同様である．

⑤ 特徴語グラフ

評価極性ごとのツイート集合における特徴語を 3 つの指標で抽出した結果閲覧できる．特徴語グラフは左から頻度法，TF-IDF，情報利得と自己相互情報量を組み合わせた指標の順で並んでおり，TOP10 まで表示されている．情報利得（以下 IG と呼ぶ）は，ある特徴が出現したか否かという情報が，クラスに関する曖昧さをどの程度減少させるかを表す尺度であり，トピックの検出などにおいて広く用いられている．ここでは，システムで収集した全ツイート集合$\mathbf{C}$に対して，入力された期間における，クエリーで入力した単語を含むツイート集合c^+を正例，それ以外のツイート集合c^-を負例とする．特徴量を求めたい語f_iの IG の計算式を以下に示す．f_i^+は語f_iが現れるツイート集合，f_i^-は語f_iが現れないツイート集合を表す.

$$IG(C|F_i) = H(C) - H(C|F_i)$$

$$H(C) = -p(c^+)\,log\,p(c^+) - p(c^-)\,log\,p(c^-)$$

$$H(C|F_i) = -p(c^+, f_i^+)\,log\,p(c^+|f_i^+) - p(c^-, f_i^+)\,log\,p(c^-|f_i^+)$$
$$-p(c^+, f_i^-)\,log\,p(c^+|f_i^-) - p(c^-, f_i^-)\,log\,p(c^-|f_i^-)$$

$$C = \{c^+, c^-\} \quad F_i = \{f_i^+, f_i^-\}$$

IG の値が高い語は，入力された期間における，クエリーで入力した単語を含むツイート集合に現れやすいか，あるいは現れにくいかのどちらかの特徴を持っていると考えられる．そこで，本研究では，IG と自己相互情報量（以下 PMI と呼ぶ）の組み合わせにより，クエリーで入力した単語とともに現れやすい特徴語の抽出を試みる．PMI は 2 つの確率変数の共起のしやすさを計る尺度である．語f_iにおける PMI の計算式を以下に示す.

$$PMI(c_0^+, f_i^+) = log\,\frac{p(c_0^+, f_i^+)}{p(c_0^+)p(f_i^+)}$$

PMI の値が正となる語f_iのうち，IG の値が上位である語をクエリー入力で取得したツイート集合における特徴語として提示する．

また，上部の「ポジ」,「ネガ」,「評価なし」のタブをクリックすることで各評価極性のツイート集合の特徴語を閲覧できる．また，上部の「1-gram」,「2-gram」,「3-gram」タブで n-gram ごとの結果を表示させることができる．特徴語グラフをクリックすることで，クリックした特徴語を含むツイートのみをタイムライン上に表示させる事ができる．

⑥ タイムラインビュー

評価極性ごとの収集したツイートを閲覧することができる．特徴語グラフと同様，上部のタブをクリックすることで評価極性の切り替えができる．表示されるツイートは，特徴語グラフで表示される単語をより多く含んでいるものほど重要なツイートであると考え，ビューの上位に表示される．

4. 評価・考察

本システムを使用して, 被験者 10 人に対して下記の問題 2 問の解答結果と，システムに関するアンケートを実施することで，開発目的を達成できているか確認した.

問題1. 昨年 12 月 14 日に行われた第 47 回衆議院議員総選挙に関して，全国的に争点となっているもの，沖縄県で争点となっているものは何か．それぞれ答えよ．（複数解答可）

問題2. 沖縄県での「辺野古への新基地移設」に対する各政党・県民の立ち位置（賛否や支持政党とその理由など）をまとめてください．（沖縄県での候補者は自民 4 人，共産・社民・維新・生活・無所属から各 1 人出馬している）

前提条件として，本システムではクエリーを入力してから分析結果を表示するまでに時間がかかるため，クエリーは「争点」,「自民党」,「共産党」,「社民党」,「維新の党」,「生活の党」,「無所属」の 7 点が入力され，検索日付範囲は選挙の投票日となる 12 月 14 日の前日である 13 日から一週間前の 6 日であ

ると仮定して，分析をあらかじめ行っておき，過去ログからクエリーを選択してもらうことで実証実験を行った．

4.1. 問題 1 の解答結果

問題 1 の解答を集計した結果の一部を表 2 に示す．

表 1:問題 1 解答結果の一部

全国		沖縄県	
争点	人数	争点	人数
集団的自衛権	10	辺野古への新基地移設	10
原発再稼働	8	米軍基地	5
アベノミクス	5	集団的自衛権	3
消費税増税	5	建白書	2
特定秘密保護法	4	アベノミクス	1
社会保障	4	安倍政権の是非	1

全国的な争点としては「集団的自衛権」や「原発再稼働」，「経済政策」などが特に多く解答されていたが，沖縄県の争点としては「辺野古への新基地移設」を全員が解答していた．これは，沖縄県での最大の争点となっていたのが「辺野古への新基地移設」であり[8]，それに関わるツイートが多く提示されていたためである．
以上のような結果から，地域ごとで分析をすることで地域の特徴的な意見の抽出が行えていると考えられる．

4.2. 問題 2 の解答結果

問題 2 に関して，実際の事実を，箇条書してまとめると以下のようなことが言える．

①. 自民党は辺野古への新基地移設を推進している
②. 共産党，社民党，生活の党，無所属の候補者は移設に対して反対派である
③. 維新の党の候補者は移設について知事選では「県民投票」を呼びかけていたが，衆院選では中止，撤回を求めている
④. 維新の党本部は辺野古への新基地移設を推進している[9]
⑤. 共産党，社民党，生活の党，無所属の候補者が沖縄の全区で当選したことから県民は基地の移設には反対で，共産党，社民党，生活の党，無所属を支持する人が多い

①，②に関しては解答者全員が正しくまとめられており，自民党と共産党，社民党，生活の党，無所属の 4 党が対立関係にあることが記入されていた．
③，④に関して，維新の党が移設に対して推進派と記入した人は 5 人，反対と記入した人は 2 人，解答がなかった人が 3 人という結果であった．維新の党は，党本部と候補者で意見が違っていたことが原因で，推進派と反対派で意見がわかれてしまったと考えられる．また，維新の党本部と候補者で意見が異なっている，というところまでまとめることできた人はいなかった．
⑤に関して，解答者全員が「県民は反対派が多く，共産党，社民党，生活の党，無所属を支持」と記入していた．理由としてあげられていたものには以下のものがある．

- 県知事選で自民が推薦していた仲井真知事が基地建設に関わる承認をし，県民の怒りをかったため，自民以外を支持する人が多い
- 歴史的経緯から基地の撤去をもとめている
- オスプレイの危険
- 県民・市民ではない人が増えることへの不満

ツイートを実際に閲覧することで，「怒りをかった」や，「危険」といった恐怖心，「不満」など県民がどのように感じているかの意見情報が読み取れていた．

4.3. アンケートと解答結果

アンケートの内容は以下の 6 問である．解答には「1.大いにそう思う」，「2.少しそう思う」，「3.どちらとも言えない」，「4.あまりそう思わない」，「5.全くそう思わない」の 5 段階で解答してもらった．ただし，質問 6. は自由記述である．

質問 1. 本システムを用いることで地域の特徴的な意見を抽出するのに役立つと思いますか
質問 2. 本システムを利用することで効率的に意見をまとめることができると思いますか
質問 3. 本システムは意見の抽出に必要な機能が揃っていると思いますか
質問 4. 本システムは意思決定支援やマーケティング，世論調査などに役立つと思いますか
質問 5. 今後，本システムを使いたいと思いますか
質問 6. 意見を読み取れる上で役に立った機能、使いづらい機能、また、あると便利だと思う機能があれば記入してください（自由記述）

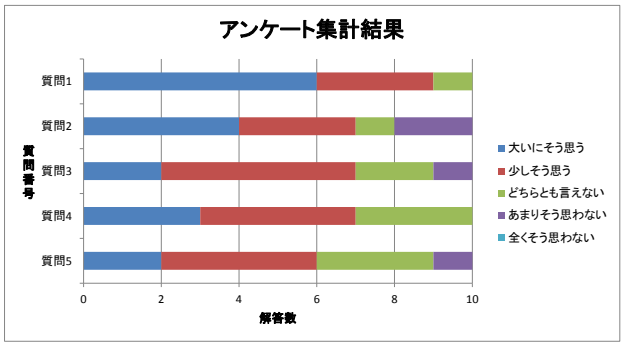

図 4:アンケート集計結果

質問 1 に対して肯定的な解答を示した人が 90%であることから、地域の特徴的な意見を抽出するのに有用なシステムの開発が達成できていると考えられる．これは問題 1 の解答結果で，全国の争点では出現していない「辺野古への新基地移設」や「建白書」などの沖縄独自の問題が争点として上げられていることや，問題 2 で県民の意見を正しく読み取れていたことからも推測できる．

質問 2,3,4,5 に対して肯定的な解答を示した人が 60～70%で質問 1 と比較すると少し減少した. 質問 4 に関しては，本実験では世論調査のためのみが対象であったため，意思決定やマーケティングにも有用であるかがわからないことが原因と考えられる．質問 2,3,5 に関しては, 自由記述にレイアウトや欲しい機能などに関する意見が多かったことからユーザビリティの低さが原因と考えられる．

4.4. 考察

本実験で，地域ごとに分析を行い，探索的にツイートを閲覧していくことで地域の特徴的な意見の抽出が可能であると示唆する結果を得た．しかし，今回出題した沖縄県の地域特性に対して，元々被験者が知識を持っていた可能性が拭えず，その場合，実験結果に有利に働いた懸念が残される．本来は，もっと認知度の低い地域特有の題材を出題するのが理想であった．しかし，今回対象にした選挙では，沖縄県以外で地域特有の争点が顕著でなく，認知度の低い地域特有の題材を多くツイートした地域を発見できなかった．今後は，そのような認知度の低い題材に対する評価実験を行う必要がある．

また，Twitter の利用者は若者が多いが，選挙での年齢別投票率は 20 代が一番低く，投票率の高い中年・高齢者層の人の意見を抽出できない可能性があり，題材によっては実際の事実とは異なる結果になる可能性がある[10]．そのため，Twitter の利用者層を考慮したシステムの利用が必要である．

5. おわりに

本稿では，Twitter から意見・評判情報を抽出する研究では十分に扱われていなかった，地域ごとの人々の意見・評判を考慮し，地域の特徴的なツイートの探索支援システムについて述べた．

また, 本システムを用いて第 47 回衆議院議員総選挙を題材に特定の地域の意見・評判情報を抽出することが可能であるか確認を行った．実際の選挙結果と比較して，地域の人々の意見を正しく抽出出来ていることを確認し，システムに関するアンケートにて開発目的を達成出来ていることを確認した．しかし，考察で示したように，被験者の元々の知識が結果に影響していた可能性もあるため，新たな題材でのシステムの評価をすることが今後の課題である．

謝辞

言語解析 Web API をご提供頂いた株式会社エクシング様に深謝します．本研究の一部は，JSPS 科研費若手研究(B) (No.25870321)の助成を受けたものです．

参考文献

[1] 選挙毎日: ネット選挙 ツイッター分析 http://senkyo.mainichi.jp/2013san/analyze/20130731.html

[2] Yahoo! Japan リアルタイム検索, http://search.yahoo.co.jp/realtime

[3] 村上奈緒，尼岡利崇，“Twitter 上で任意の検索語句に対するネガポジ度を判定し可視化するアプリケーションの開発と研究”，エンタテインメントコンピューティングシンポジウム (2014)

[4] Itokawa, S.,Shiramatsu, S., Ozono, T., and Shintani, T., "Estimating Feature Terms for Supporting Exploratory Browsing of Twitter Timelines," In *Proc. of IIAI-AAI 2013*, pp. 62-67 (2013)

[5] 北本 朝展, 相良 毅, 有川 正俊, “GeoNLP：自然言語文を対象とした高度なジオタキングに向けて”, In *Proc. of CSIS Days 2011*, No. D10 (2011)

[6] 言語解析 WebAPI, 株式会社エクシング, http://bigdata.joysound.com/about.html

[7] MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[8] 琉球新報，“衆院選(衆議院議員選挙)”, http://ryukyushimpo.jp/news/storyid-235313-storytopic-125.html

[9] 維新の党マニフェスト, https://ishinnotoh.jp/election/shugiin/201412/pdf/manifest.pdf

[10] 衆議院議員総選挙における年代別投票率の推移, http://www.soumu.go.jp/senkyo/senkyo_s/news/sonota/nendaibetu/