

地域活性化のための意味表現学習を用いた学生のツイート分析

櫻田直行 浅野有治 芥子育雄

福井工業大学 〒910-8505 福井県福井市学園3丁目6-1

E-mail: {e16049ns, e16001ya, keshi}@fukui-ut.ac.jp

あらまし 筆者らは、Twitter を用いたデータの分析による地域活性化を目的とし、その焦点を福井の学生に集中することでイベント情報や学生の行動などを活用することを考えた。データ分析には、264 種類の特徴単語で単語やパラグラフの意味を代表する意味表現学習を用いた。学生約 100 名のツイート 5 万件を特徴単語ごとに重みが大きい順に分類し、閾値以上の特徴単語の重みを持つデータを分析した結果、有益な情報がフィルタリングできることを確認した。今後、学生の関心が集まるトピックを自動的にまとめるアプリケーションを作成することにより付加価値を与え、データ量を増大させることを目指す。

キーワード 意味表現学習, Twitter, 地域活性化, 学生, SNS

1. はじめに

近年、福井県では地域活性化を目指す動きが大きくなりつつある。例えば、地域イベント助成事業や公共スポーツ施設等活性化助成事業、宝くじスポーツフェアなどの支援を行うことで、より地域への密着度を高める動きなどがある。これらの主な内容は子供のスポーツ促進や小規模イベントの支援といったものに重点が置かれている[5]。一方で学生向けのイベントも行われているが、認知度が低く、学生の意見の取入れや学生主催のイベント情報の拡散性が低いことが問題となっている。

そこで本研究では、学生の意見やイベント情報などを Twitter で収集し、アプリケーション化することを目指す。Twitter を利用することで、学生の直接的な意見や感想、そのイベントへの期待値を取り入れることが可能だと考えた。また、アプリケーション化することでよりイベントの認知度を高め、ローカルイベントへの参加者を増加することが可能だと考えた。



図 1 本研究で使用したアカウント

2. 本研究の対象と方法

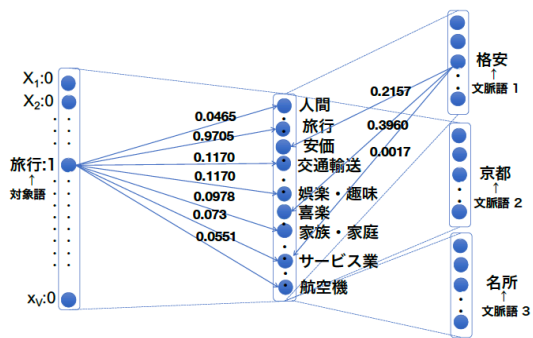
本研究のデータ収集の際に使用したアカウントを図 1 に示す。解析した Twitter のデータの対象者は、福井県の大学に在籍している学生 113 名である。また、2018 年 12 月 1 日時点での 57,961 ツイートを収集した。これらのツイートはリツイートやリプライを除いたものである。本研究では、ツイッターアカウントを作成し、人手によって学生のみを選別した。基準として、プロフィールに大学名もしくは学部が入力されているアカウント、または Twitter 外で福井県の大学に在学していることが確認できる学生のみを選別した。

選択したアカウント 1 人あたり直近の 3000 件のツイート（リツイートやリプライを除くと、各学生のツイート数は 3000 件に達していないため、全ツイートが対象）を Twitter API を用いて収集した。

3. 意味表現学習

ニューラルネットワークを利用した単語やパラグラフの分散表現学習[3, 4]は、学習結果の解釈性の課題がある。筆者らは、分散表現学習に専門家が構築した単語意味ベクトル辞書（以降、辞書と呼ぶ）を反映する意味表現学習を提案した[1]。辞書に登録されている約 2 万語（基本単語と呼ぶ）は、264 種類の特徴単語との関係により意味を表現する。図 2 に示す通り、ニューラルネットワークの隠れノードに特定の意味（特徴単語）を割り当て、辞書を元に基本単語にシードベクトル（初期の重み）を設定することにより、学習後も重みの大きな隠れノードの意味が維持されることを示した[1]。

意味表現学習応用システム構築の流れを図 3 に示す[2]。最初に基本単語のシードベクトルを元に、約 120 万記事の日本語 Wikipedia コーパスを用いて、単語ベクトルの学習[3]を行う。次に分析対象となる Twitter コーパスを用いて、パラグラフベクトルの学習[4]を行



う。

図 2 辞書を反映したシードベクトルのイメージ

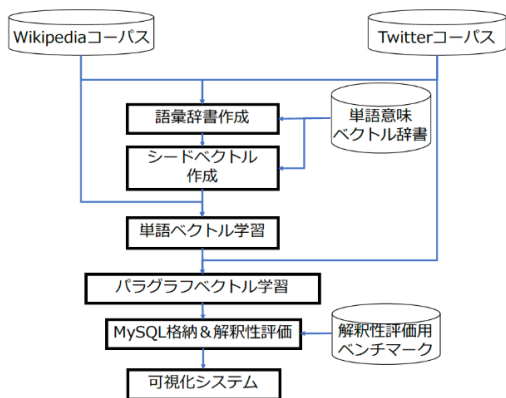


図 3 意味表現学習応用システム構築の流れ

シードベクトルには Wikipedia コーパスで学習した単語ベクトルを用いる。パラグラフベクトルを学習したツイートの中から特定の特徴単語の重みを降順にソートし、上位 N 件のツイートを高速に抽出できるように各ツイートのパラグラフベクトルを関係データベースの MySQL に格納する。MySQL には、Twitter ID、ツイートを代表する代表ワード、264 種類の特徴単語の重みを格納する。

4. 分析結果

意味表現学習を用いて特徴単語ごとに閾値以上の重みをもつデータ、上位約 1000 件を MySQL から抽出した。データの分析を行った結果、特定の特徴単語の重みの大きさにより、行動（特徴単語：人間、人名）や関心のあるスポーツ（特徴単語：スポーツ）、ローカルイベント（特徴単語：社会活動）などの有益な情報

（以降、トピックと呼ぶ）がフィルタリングできることを確認した。

トピックごとの適合率を表 1 に示す。特徴単語 1 つにつき上位 50 件を調べた。“行動”は特徴単語 2 つを対象とした分析のため各 50 件を合算した 100 件とした。そこから広告などを除いたものを対象データ数とし、該当データ数はトピックに関連の大きいデータを人手によって選択した数である。また、適合率は、（該当データ数/対象データ数）で求めた。

イベント情報、特にローカルイベントは、大々的に告知されていないものも多くみられた。一方でスポーツでは学生が関心のあるスポーツの傾向は分かるが、同世代の有名選手がいるにもかかわらず、そのスポーツのツイートが上位にあがることはなく、競技人口の多い野球やサッカーが上位を占める結果となった。また、行動では、学生の学内外での活動内容、趣味や娯楽といった勉強以外にどのような活動をしているかが見て取れた。さらに旅行では、学生の旅行先や観光場所の情報を収集できることが分かった。

5. 関連研究

ツイッターを用いた分析をする場合、収集する対象の選定とツイート収集の効率化が課題となる。

田原[6]らは、特定地域のユーザを検索する研究をしている。特定地域の情報を日常的に発信するユーザは、その地域特有の場所・日常的な話題などを示す語を多く発信していると考えられる。そこで地域特徴語を Twitter から抽出し、より多くのツイートするユーザをその地域情報の発信者とした。本研究では、ユーザの選定は人手によって行ったため、この研究を活用することが考えられる。

長島[7]らは、ツイートの選別を行い、特定の地域の店舗の口コミのツイートのみを抽出する研究をしている。位置情報や地域名を含まないが店舗名を含むツイートを収集するために地域ユーザを用いることで口コミツイート数を増やすことに成功している。地域ユーザはツイプロ（プロフィール検索サービス）を用いて収集しており、本研究に取り入れることでさらに学生数、ツイート数を増やすことが期待できる。

中川[8]らは、特定地域から発信されたツイートの効率的な収集方法の研究をしている。この研究は対象の

表 1 トピックごとの適合率

トピック	特徴単語	対象データ数	該当データ数	適合率 (%)
行動	人間, 人名	94	65	69.1
旅行	旅行	50	42	84.0
イベント情報	社会活動	47	31	65.9
スポーツ	スポーツ	50	50	100.0
数学	数学	48	20	41.7
通信	通信	47	26	55.3

範囲が広い場合、フォローの制限や対象外の人物が含まれることを防ぐことを目標としており、探索アルゴリズムや発信地推定などを用いることで実現している。本研究では福井の学生という狭い範囲での対象となっているため、フォローの制限には問題がないと考えられるが、卒業生や県外の学生などを対象としないように注意が必要である。発信地推定や探索アルゴリズムを用いることも考慮する必要がある。

6. まとめ

今後の課題として、現在のデータ量では特徴単語ごとに分析する際、重みの低いものが上位に来ることが多い。また、ゲームや他サイトの告知といった不純物が上位にくることがあるため、データの選別とデータ母数の増加をする必要がある。また、卒業した学生をデータの対象外にすることも課題のひとつに挙げられる。

本研究の活用方法として、分析したデータをアプリケーション化し、まとめることによって福井で行われるローカルイベント情報や学生の行動を分析することが可能となると考えている。図4にアプリケーション案を示す。また、行動面を見ることで学生の活動場所が具体的にわかるため、より効果的な告知をすることが可能になる。今後、福井でイベントを行う際、学生へのイベントの浸透性や関心などを把握するために利用可能と考えている。

参 考 文 献

- [1] Ikuo Keshi, Yu Suzuki, Koichiro Yoshino, Satoshi Nakamura, "Semantically Readable Distributed Representation Learning and Its Expandability Using a Word Semantic Vector Dictionary," IEICE TRANSACTIONS on Information and Systems, Vol.E101-D, No.4, pp.1066-1078, Apr. 2018.
- [2] 芥子育雄, 松田義貴, 鈴木優, 吉野幸一郎, 中村哲, 意味表現学習におけるツイート分散表現の解釈性評価と可視化の提案, DEIM Forum 2018 B1-1, Mar. 2018.
- [3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," Proc. of NIPS, pp.3111-3119, 2013.
- [4] Q. Le, T. Mikolov, "Distributed Representations of Sentences and Documents," Proc. of ICML, pp.1188-1196, 2014.
- [5] 地域や地域コミュニティの活性化への支援, 福井県ホームページ, <http://www.pref.fukui.lg.jp/doc/sityousinkou/komyunitysien.html>, 2012.
- [6] 田原琢士, 馬強, Twitter から有益な日常情報を発見するための特徴語による地域ユーザの検索, DEIM Forum 2014 B3-3, Mar. 2014.
- [7] 長島里奈, 関洋平, 猪圭, 地域ユーザに着目したロコミツイート収集手法の提案, DEIM Forum 2016 B4-3, Mar. 2016.
- [8] 中川真史, 山口祐人, 北川博之, フォローを用いた特定地域から発信されたツイートの効率的な収集, DEIM Forum 2018 B3-3, Mar. 2018.

サイト名		まとめ
アイコン ID : aaaaaa プロフィール 使用された自分のツイート数 :	行動 学校帰りに とカラオケ?? とおしゃべりはホント笑いが止まらんかったーお互いにコンタクト忘れて裸眼辛すぎたけど楽しかったから ○?いつかは裸眼美女になるーネ? ケータイ海で水没したんでしばらくline等連絡つきません笑死ぬほど楽しかったのにいいいい!! わざわざ愛知から来てもらったのにトラブルメーカーでごめんね?また遊ぼうね! 今度は私がまた豊橋にお邪魔します 女子4人でキング来て隣の女子会が2人前頼むのを躊躇してる中タンだけで19人前食べきった?@ 昨日幼馴染連中とあそびただけなんだろうこのおじいちゃんおばあちゃんの旅行に付き添いで来た息子感笑笑 今日は放課後に体育館借りて3年野球部寮生 + やで3時間バスケットで明日の球技大会はみんな筋肉痛やな?その後3年生寮生野球部サッカー部でカルビ大将*俺はもう引退してから寮じゃないのに呼んでくれるとこでみんなの愛... 市民音楽祭わづりんちょ楽しかった~♡けども!!本番後の練習はキツついぜ~~~~~!!!!!!笑はあ...本番までに体力つけなきゃ? 部室で2時間も寝てるの意味わからんのやけど..... さんの寝言エンター解放ダァ!	
	トピック	・行動 ・旅行 ・スポーツ ・通信 ・イベント 関連性 ~100 ▼ 性別 男 ○ 女 ○

図4 アプリケーション案