5Q-07

SNS を対象としたトピックモデルによる時系列話題遷移抽出方式

田澤 紗彩† 岡田 龍太郎† 中西 崇文†
武蔵野大学データサイエンス学部データサイエンス学科†
E-mail: s1922066@stu.musashino-u.ac.jp, {ryoutaro.okada, takafumi.nakanishi}@ds.musashino-u.ac.jp

1. はじめに

近年，SNS(Social Networking Service)やブログなどの新しいメディアを活用しているユーザが増加しており，これらの環境下でユーザ自らの思想や感情を発信し，共有することが多くなってきている．一方で，2020 年は，新型コロナウイルス感染症の流行により，未曾有の危機を迎えており，感染者数と死亡者数が増え続ける中，未知なる新型コロナウイルスの感染に対して，人々がどのように考え感じ行動したのかを分析することは，感染症対策を決める上でも一つの指標となりうる．特に，人々の考えや感じ方，行動の変化が，SNS(Social Networking Service)やブログなどの新しいメディアを通じて収集，抽出することができれば，未曾有の危機の中の人々の生の声を客観的に知ることができると考えられる．

新型コロナウイルス感染症の状況は時事刻々と変化しており，この変化に応じて，人々の考えや感じ方，行動が時事刻々と変化していると考えられる．そのことから，新型コロナウイルス感染症に関する SNS の投稿を日々収集し，その内容の時系列変化を捉えることができれば，人々の生活の一端を観察することが可能となり，感染症対策を考慮する上での一助となると考えられる．

本稿では，SNS を対象としたトピックモデルによる時系列話題遷移抽出方式について示す．本方式は，Twitter のツイートデータを対象として，ある話題に関連するツイートデータを時系列に収集し，そのツイートデータのトピックを Latent Dirichlet Allocation(LDA)[1]を用いて日ごとに抽出することにより時系列における話題の変遷を示す．本方式は，Twitter 上で呟かれる特定の話題についての話題の変遷を示すことで，その話題の関心や注目度を俯瞰的に観察することが可能である．本稿では，新型コロナウイルスに関するツイートデータを収集し，新型コロナウイルスの第 1 波流行時における話題の変遷について示す．

Time-Series Topic Transition Extraction by Utilizing Topic Model for SNS
Saya Tazawa†, Ryotaro Okada†, Takaumi Nakanishi†
† Musashino University, Department of Data Science

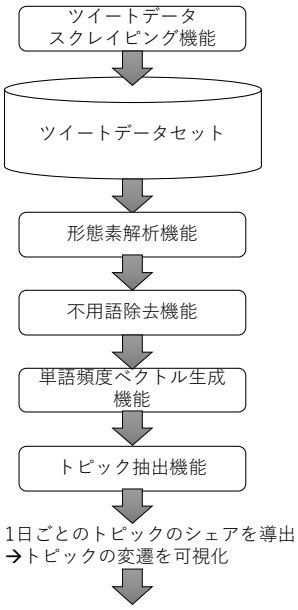

図 1. 本方式の全体像

2. 関連研究

鳥海ら[2]は，Twitter 上で新型コロナウイルスに関係する単語を含んだツイートを収集し，ML-Ask を用いた感情分析を用いて，ツイートデータ群からその感情の時系列変化を抽出している．

本稿では，ツイートデータのトピックを Latent Dirichlet Allocation(LDA)[1]を用いて日ごとに抽出することにより時系列における話題の変遷を示すことを試みている．

3. トピックモデルの時系列話題遷移抽出方式

3.1 全体像

本提案方式の全体像を図 1 に示す．本方式は，ツイートデータスクレイピング機能，形態素解析機能，不用語除去機能，単語頻度ベクトル生成機能，トピック抽出機能からなる．

3.2 データセット

本稿で用いるデータセットは，3.3 節で述べるツイートデータスクレイピング機能で収集する．具体的には，「コロナ」という単語を含むツイートで，かつ，「いいね」と「リツイート」の両方が 100 件以上のツイートを対象に収集をした．その結果，2020 年 4 月 1 日から 2020 年 6 月 30 日

までで，件のツイートデータを収集した．

3.3 ツイートデータスクレイピング機能

TwitterAPI を通じて，キーワード，いいねの件数，リツイートの件数の条件を指定して，その条件に合致したツイートデータを取得できる機能である．

3.4 形態素解析機能

3.3 節で収集した全てのツイートデータを対象として形態素解析を行い，名詞，動詞，形容詞のみをそのツイートの重要な単語として抽出を行う．

3.5 不用語除去機能

ストップワードリストに則り，3.4 節で抽出された単語から不要語を削除する．ストップワードには，URL の断片，こそあど言葉，「コロナウイルス」，「新型コロナウイルス感染症」，ニュースサイトの名称を挙げている．

3.6 単語頻度ベクトル生成機能

3.5 節で残った単語群を用いて，1 日ごとに単語頻度をカウントした上でベクトル化をする．これにより，1 日ごとにどのような単語が頻出したツイートがあったかを確認することが可能となる．

3.7 トピック抽出機能

3.6 節で得られたそれぞれの単語頻度行列を用いて Latent Dirichlet Allocation(LDA)[1]を用いてトピック抽出を行う．これにより，ツイートを収集している期間でどのような話題があったのかを明示的に分けることが可能である．また，1 日ごとにそれらのトピックがどのようなシェアがあったのかを導出する．これにより，話題の遷移をトピックのシェアで判断することが可能となる．

4. 実装

3 節で提案した方式を実装し実験を行った．抽出されたトピックとトピック語の上位 15 件を表 1 に示す．網掛けなっているトピック語は，15 件までの他のトピックに存在していないユニークな語である．トピック名は筆者らが設定した．また，5 月のツイートの分析によるトピック推移を図 2 に示す．

5 月には，19 日にアメリカのトランプ大統領が WHO 脱退を示唆するニュースが報道され，世界情勢にあたるトピック 3 のシェアが大きくなっている．このように，トピックとそのシェアを見ることで情勢を把握することが可能となる．

5. おわりに

本稿では，SNS を対象としたトピックモデルによる時系列話題遷移抽出方式について示した．本方式は，Twitter 上で呟かれる特定の話題についての話題の変遷を示すことで，その話題の関心や注目度を俯瞰的に観察することが可能である．本稿では，新型コロナウイルスに関するツイートデータを収集し，新型コロナウイルスの第 1 波流行時における話題の変遷について示した．

今後の課題は，時系列テキストデータを対象としたトピック抽出機能の実現と適用，新型コロナウイルスのゲノムデータ[3]を用いたゲノム変異状況との連動，他の重要事象における本方式を用いた分析の適用が挙げられる．

表 1. トピック語とトピック名

トピック 0 国民の生活	トピック 1 国内政治	トピック 2 東京の情勢	トピック 3 世界情勢
憲法	言う	言う	感染者
出る	やる	患者	世界
対応	マスク	見る	8%
受ける	中国	何	政府
必要	憲法	わかる	検査
増える	問題	出る	コロナ禍
言う	コロナ禍	コロナ禍	中国
支援	みんな	本日	見る
危機	検査	考える	WHO
生活	見る	病院	補償
状況	情報	私	自粛
社会	安倍首相	可能性	支援
緊急事態条項	国	検査	話
病院	補償	国会	自分
検査	中止	東京	陽性

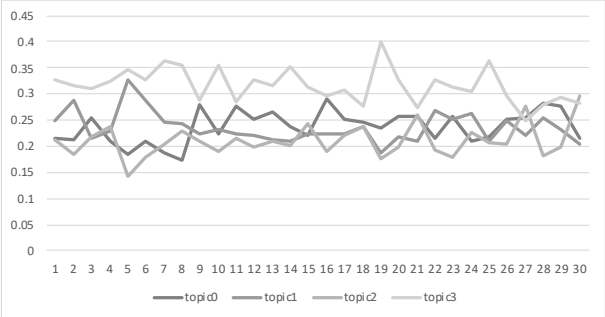

図 2. 5 月のツイートのトピック推移

参考文献

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research, 3(Jan), pp. 993-1022, 2003.
[2] 鳥海 不二夫，榊 剛史，吉田 光男，ソーシャルメディアを用いた新型コロナ禍における感情変化の分析，人工知能学会論文誌，35(4)，F-K45_1-7, 2020.
[3] GISAID, https://www.gisaid.org/