4B-03

ツイートデータによるコロナニュースの発信システム

山田　実俊†　　今西　規‡
東海大学医学部†　　東海大学医学部‡

1. はじめに

新型コロナウイルス感染症が世界中で猛威を振るっている。新規感染者数や新変異株の出現などの情報には多数が感心を持つため、迅速かつ正確な情報収集が望まれる。ソーシャルメディアには速報性があり、読者の反応によりニュースの重要性を評価できる点で既存のメディアよりも優れている。

そこでわれわれは、ソーシャルメディアのひとつである Twitter を利用してコロナ関連のニュースを発見し配信するシステムを開発した。「コロナ」という単語を含むツイートから全単語の出現頻度の推移を解析して注目単語を抽出し、共起ネットワークを用いてツイート内容を俯瞰的に可視化した。この結果をリアルタイムに更新することで、ニュース価値の高い情報を発見・発信するシステムを作成した。

2. ツイートデータ

文章の中に「コロナ」を含むツイートを 1 分ごとに 1,000 件ずつ収集を行った。ツイートの収集には統計ソフト R の rtweet パッケージを利用した。収集したツイート文を閲覧したとき、企業の広告やキャンペーンに関するツイートが多く見られた。まず分析を開始する前に、ツイートデータから広告やキャンペーンに関する単語や、「アカウント」、「フォロー」、「フォロワー」、「リツイート」、「いいね」を含むツイート、Twitter bot と思われるアカウント（アカウント名に「bot」が含まれる）のツイートなどを分析対象から除外した。また、ツイート文の全角英数字を半角に統一し、URL やリプライ先のアカウント名などの削除も行った。

収集したツイートデータについて形態素解析を行った。形態素解析のソフトは MeCab を使用し、R の RMeCab パッケージを利用した。このとき、新型コロナウイルス感染症に関する新しい用語や類義語については、自作の辞書を作成し適用させた。例えば「テレワーク・在宅ワーク・在宅勤務・自宅ワーク・自宅勤務・リモートワーク」は全て「テレワーク」に置換した。

3. 注目単語の抽出

出現頻度が急上昇している単語はその時間帯に注目されていると考えられる。時系列データのバースト性については 2 群の比較を用いた研究[1]もあるが、本研究は個々の単語のバースト性を考慮した方法を提案する。

形態素解析で得られた単語（名詞・形容詞）の出現頻度を 5 分ごとに集計した。2 時間分のデータ（例:15 時 00 分～17 時 00 分）に対して、単語 A の始点の時間（例の場合は 15 時 00 分）から t 分後の注目度($HT_t(A)$)を以下のように算出した。

$$HT_t(A) = \frac{Fr_t(A)}{Fr_0(A)}$$

ここで、$Fr_t(A)$は単語 A の t 分後の出現頻度であり、$Fr_0(A)$は単語 A の始点の時間の出現頻度である。また$Fr_0(A) < 10$のとき、注目度が高くなってしまうため、$Fr_0(A) = 10$として算出している。そして$HT_t(A) > 5$のとき単語 A は注目されたとみなし、注目単語としてリストアップした。

図 1 左上は 2021 年 12 月 1 日の 15 時 00 分から 17 時 00 分に抽出された注目単語の出現頻度の推移を示す。x 軸は時間帯、y 軸は 5 分ごとの出現頻度を表す。折れ線グラフ上の単語の位置は、その単語の出現頻度が最も多かった時間帯と出現頻度を表し、他の単語と重ならないように表示している。抽出された注目単語の一覧を図の下に記載している。また単語の枠の色は、青色が報道関連のアカウントによる投稿、赤色がそれ以外のアカウントによる投稿である。

図 1 左上の例では注目単語が 44 単語抽出され、主に 3 か所の時間帯で単語が集まっている。16 時 50 分頃の「感染」、「確認」、「東京都」、「21 人」などの注目単語は、毎日 16 時 45 分頃に東京都の本日の感染者数が公表されており、そのニュース記事をリツイートするユーザーが多くいることがわかる。15 時 50 分頃の「国内」、「2 例目」、「ペルー」、「新変異株」などの注目単語は、新変異株であるオミクロン株の国内 2 例目の感染者が確認されたニュースが話題となり、17 時 00 分になっても注目度があまり下がっていないことがわかる。15 時 10 分頃は、あるアーティストがコロナ禍でのライブツアーについてコメントし話題になったが、オミクロン株のニュースのように長く続く話題とはならなかったことがわかる。

4. 注目単語の共起ネットワーク

実際にどのようなツイート内容が注目されて

COVID-19 news transmission system based on tweet data
† Sanetoshi Yamada, Tokai University
‡ Tadashi Imanishi, Tokai University

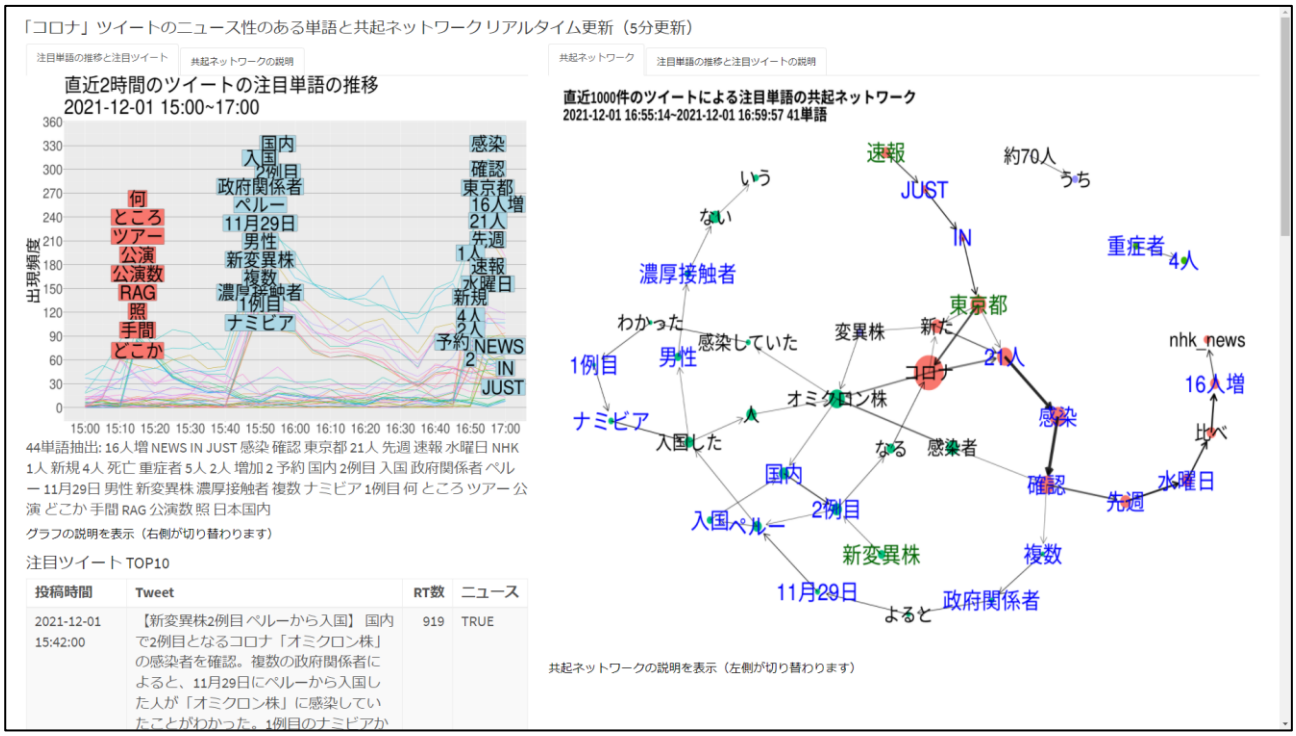

図 1. 「コロナ」を含むツイートの注目単語の推移と共起ネットワークを表示する可視化システム

いるかを表現する方法として、注目単語の共起ネットワークを作成した。この共起ネットワークは災害時のツイート分析の研究成果[2][3]を改良して作成した。収集した約 1,000 件のツイートの形態素解析で得られた単語（名詞・形容詞・動詞）をツイート文中の出現順に並べ、連続する 2 つの単語を共起として出現頻度（$Fr(A \rightarrow B)$）を集計した。このとき動詞については、後ろに続く助詞や助動詞も 1 つの単語として処理をする改良を行った。そして出現頻度上位 50 組の共起を抽出し共起ネットワークを作成した。

図 1 右は 2021 年 12 月 1 日の 17 時 00 分に収集した約 1,000 件のツイートから作成した注目単語の共起ネットワークを示す。矢印の太さは共起の出現頻度、矢印の濃さは共起率（$CR(A \rightarrow B|A) = Fr(A \rightarrow B)/Fr(A)$）を表している。円の大きさは単語の出現頻度を表し、文字の色はニュース記事（青）、それ以外（赤）、ツイート文の始まり（緑）に多く使われていた単語を表している。また文字の背景の円の色はクラスター分析による単語グループを表し、同じ背景色の単語を矢印に沿って読むことで元のツイート文を連想できる。

図 1 右の例では、**3 章**で紹介した東京都の感染者数とオミクロン株のニュースが話題となっており、ツイートごとの表現の多様性をある程度吸収しつつ、約 1,000 件のツイート文を要約して図示することに成功した。

5. まとめ

リツイート数が多くても実際に話題になった時間帯にはずれが生じるため、図 1 のような注目単語の推移と共起ネットワークを同時に表示し、話題になった時間帯を把握できる可視化システムを作成した。さらにリアルタイムに更新（5 分ごと）することで、ニュース価値の高い情報を発見・発信することに成功した。最終的には Twitter を利用していない人でも Twitter からコロナ関連のニュースを取得できるように、リアルタイムな地域への感染症関連情報の提供サイト (http://covid-map.bmi-tokai.jp/)での公開を目指す。

謝辞

本研究は、東海大学連合後援会研究助成金（2021 年）の助成を受けて実施した。

参考文献

[1] J. Kleinberg. “Bursty and Hierarchical Structure in Streams”, Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[2] S. Yamada, K. Utsu, K. Cho and O. Uchida. “Analysis and Visualization of Attention Area of Tweets During Disasters,” 2019 The 6th International Conference on Information and Communication Technologies for Disaster Management.
[3] S. Yamada, K. Utsu and O. Uchida. “Visualization of Tweets and Related Images Posted During Disasters,” 5th IFIP WG 5.15 International Conference, ITDRR 2020.