# <調査報告> Webスクレイピングを利用した 医薬関連 情報 の 収集 と入手データの活用



| | |
|---|---|
| 著者 | 松延 千春, 石井 起弥, 井上 寛, 白谷 智宣 |
| 雑誌名 | 第一薬科大学研究年報 |
| 号 | 38 |
| ページ | 51-64 |
| 発行年 | 2022-03 |
| URL | http://id.nii.ac.jp/1154/00000073/ |

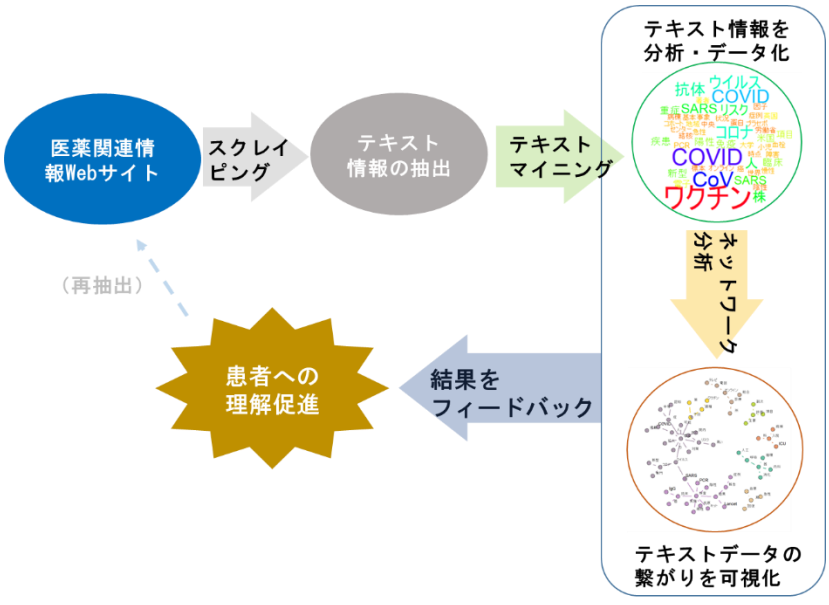
調　査　報　告

# Web スクレイピングを利用した医薬関連情報の収集と入手データの活用
# Collecting of drug-related information using web scraping and utilization of the obtained data

松延　千春[a]，石井　起弥[b]，井上　寛[b]，白谷　智宣*,[a]
Chiharu Matsunobu[a], Ishii Tatsuya[b], Hiroshi Inoue[b], Tomonori Shiratani*,[a]

[a]第一薬科大学　薬学部　薬学教育推進センター
[b]九州産業大学　経済学部
[a]Center for advancing Pharmaceutical Education, Faculty of Pharmaceutical Sciences, Daiichi University of Pharmacy
[b]Faculty of Economics, Kyushu Sangyo University

薬剤師は常に最新の医薬関連情報を迅速かつ正確に入手するために、正しい情報収集の方法を習得し、情報を整理し活用する能力を身につける必要がある。現代の情報収集においてインターネットは最も身近な媒体であり、ここから迅速に、正確な情報を入手することは必要不可欠な能力の一つになってきている。今回、プログラミングにより大量のテキストデータを少ない手間と時間で抽出できる Web スクレイピングを行い、医薬関連情報を簡便に得られるかを検討し、さらに得られたテキスト情報の活用法として、大量のテキストデータから必要となる情報を抽出するテキストマイニングによる出現頻度解析、さらに共起ネットワーク解析を実践しデータの可視化を行った。本手法は、インターネットという膨大な情報源から必要なデータを抽出し、活用するためのツールとして有効であることが示唆された。



## はじめに

薬剤師法第 1 条には薬剤師の任務として、「薬剤師は、調剤、医薬品の供給その他薬事衛生をつかさどることによって、公衆衛生の向上及び増進に寄与し、もって国民の健康な生活を確保するものとする。」と記載されている。それを果たすために、薬剤師は常に最新の医薬品等の情報を迅速かつ正確に入手するために、正しい情報収集の方法を習得し、情報を整理し活用する能力を身につける必要がある。また、医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律第 1 条の 6 には、「国民は、医薬品等を適正に使用するとともに、これらの有効性及び安全性に関する知識と理解を深めるよう努めなければならない。」とあることから、国民には医薬品等の情報を入手し、正しく理解することが求められている。

医薬品等の情報を収集するための媒体には様々なものがあるが、現代の情報収集においてインターネットは最も身近な媒体となっている。総務省、令和 2 年通信利用動向調査報告書によると、国民のインターネット利用者の割合は、平成 21 年から平成 30 年までは 8 割程度でほぼ横ばいとなっており、令和 2 年は 83.4%となっている [1)]。また、インターネット利用者の利用頻度をみると、「毎日少なくとも 1 回は利用」の割合が 81.9%となっており、さらにインターネットで利用した機能・サービスと目的・用途をみると、「情報検索」が 72.7%、「ソーシャルネットワーキングサービスの利用」が 70.2%であることから [1)]、情報収集にインターネットが十分活用されていることが想定できる。しかし、インターネットは、最新の情報を地域差なく迅速に入手することが可能であるが、情報量が膨大であるため、インターネット上の情報を効率よく、かつ正確に検索することが必要になる。そこで、近年、Web サイトから大量の情報を自動的に抽出するコンピュータソフトウェア技術である、Web スクレイピング [2)]と呼ばれる手法が活用されている。Web スクレイピングとは、インターネット上に存在する膨大なデータ群の中から、特定の情報だけを自動で抜き出し、加工する技術のことである [3, 4)]。近年、Python 等のプログラミング言語では、Web スクレイピングを行う上で重要なライブラリが増えてきている。Web スクレイピングで得られたデータの多くは HTML 形式の非構造化データで、これをスプレッドシートなどに変換することで、さまざまなアプリケーションに利用することが可能になる。そこで、本研究では、特定の Web ページから医薬関連の膨大なテキストデータを Web スクレイピングにより抽出した後、得られたテキストデータについて、大量のテキストデータから必要となる情報を抽出するテキストマイニング [5, 6, 7)]、さらにテキスト同士のつながりやネットワークを見るためのネットワーク分析 [8, 9)]を行い、医薬関連情報の収集ツールとして Web スクレイピングが利用可能であるかについて検討した。

## 方法

本研究でのデータ収集は、日経メディカルの 2020 および 2021 年（12 月 13 日まで）の記事のテキストデータを Python の Web スクレイピングライブラリ「newspaper3k」[10)] を用いて収集した。テキストデータに関して、日経メディカルの 2020 年 1 月～2021

年 12 月 13 日までの有料会員限定の記事を除くすべての記事についてのテキストデータを収集した。文字数は約 65 万文字である。「newspaper3k」とは、HTML を巡回し、例えばタイトルだけを抜き出す、本文のみを抜き出すといった作業を行うための一般的に用いられているライブラリの一つである。「newspaper3k」の特徴は、一つの記事から情報を抜き出すのではなく、ニュースサイトのトップページに表示される記事を上から順番に巡回することを得意とし、複数に渡るページからデータを収集する。さらに、HTML のタグごとにデータを抽出できるため、Web サービスが更新された場合や、巡回先の Web サイトが別のリンクへと繋がっていたりする場合であっても、そのままのプログラムでデータ抽出が可能となることである。多くのニュースサイトがそうであるように、日経メディカルにおいても別のリンクへと誘導されるケースがあり、プログラムがうまく動作しない可能性を排除できる。また、これによって、アクセスする Web サイトが違えば、HTML の構成も全く違うため、同じプログラムをそのまま流用することが困難という問題を解消している。

本研究では「日経メディカル」にアクセスし「newspaper3k」を使い、毎月のアーカイブから記事をスクレイピングすると同時に xlsx ファイルに変換するプログラムも用い、その後のデータ分析で使える状態へ加工した。なお、「日経メディカル」の利用規約では、これらの行為を禁止する記述は存在しないことを確認した上で行っている。また、サーバへの負荷を考慮するために、サイトへのアクセスを 1 秒おきに制限し、意図しない DoS 攻撃により相手側のサーバの機能を停止させるような処理を行うことのないように time モジュールの sleep 関数を用いて対策を行っている。

収集したデータの解析に関しては、プログラミング言語「R」を用い、テキストマイニング・ネットワーク分析を行った。具体的には、形態素解析に基づくテキストのデータ化を行うツールとして「Mecab」を用い [7, 11]、１単語・２単語１節・３単語１節・ワードクラウド等の頻度解析を実行した。さらに、アルゴリズム（性質）の異なる２種類のネットワーク分析手法「Link Community」[11, 12] と「Overlapping Cluster Generator」[11, 13] を用い、単語のつながりを分析した。この２つのネットワーク分析手法の特徴は、１単語が分類されるクラスターは単一ではなく、１単語に対して複数のクラスターに分類される点であり、これにより、テキストデータの中で、重要なハブとなる単語を把握することができ、ハブとなる単語に関連するネットワークを見ることができるところにある。また、２つのネットワーク分析手法の違いは、「Link Community」では、大きな community とネットワークを見ることができ、「Overlapping Cluster Generator」では、より詳細な community とネットワークを見ることができるところである。

## 結果

2020 年および 2021 年の１年間トータルの Web ページに関して Web スクレイピングにより抽出したテキストファイルについて、テキストマイニングおよびネットワーク分析に基づいた結果を示す。

1. 1 単語のみの頻度解析およびワードクラウド

頻度解析とは、大量のテキストからどのワードが多く頻出しているかを調査するもので、カテゴライズ分析とも呼ばれる。まず、1 単語のみの頻度解析結果（Fig. 1）およびより視覚的にテキストデータを捉えることができるワードクラウド（Fig. 2）の結果を示す。

2020 年は、新型コロナウィルス感染症が初めて世界中にまん延したことにより、当然ながらそれに関するワードである COVID やウィルス、コロナ、新型などが上位を占める結果が得られた。2021 年度も終息を迎えていないことから、引き続き新型コロナウィルス感染症に関するワードが上位に来ているが、ワクチンや抗体などのワードが代わりに上位に来ていることが頻度解析やワードクラウドからわかる。

2. ２単語１節、３単語１節における頻度解析

1 単語での頻度解析では、ダイレクトに Web 上で用いられた単語総数がみてとれるが、その単語の前後のつながりがわからないため、どのような状況でその単語が用いられたかがわかりにくい。また、長い単語になると抽出そのものが困難になる。そこで、２単語１節（Fig. 3）、３単語１節（Fig. 4）における頻度解析を行ったので、それぞれの結果を示す。

2020 年の２単語１節の頻度解析（Fig. 3、上）から、1 単語の頻度解析で上位にあった「ウィルス」は、コロナや感染などのワードとセットになっていることがわかる。さらに、「感染」というワードは、１単語の頻度解析では上位３０ワードにはなかったが、２単語１節の頻度解析のデータではウィルスや拡大、院内、リスクなどとセットで用いられていることがわかる。2021 年になると、1 単語での頻度解析で上位であったワクチンに関連する、有効・性や臨床・試験、初回・接種、予防・効果などのワードがみられる（Fig. 3、下）。

興味深いのは、2021 年の３単語１節における頻度解析で見られる、診療・報酬・改定である。2022 年の診療報酬改定についてのトピックが記載されていたことがこれからもわかる（Fig. 4、下）。

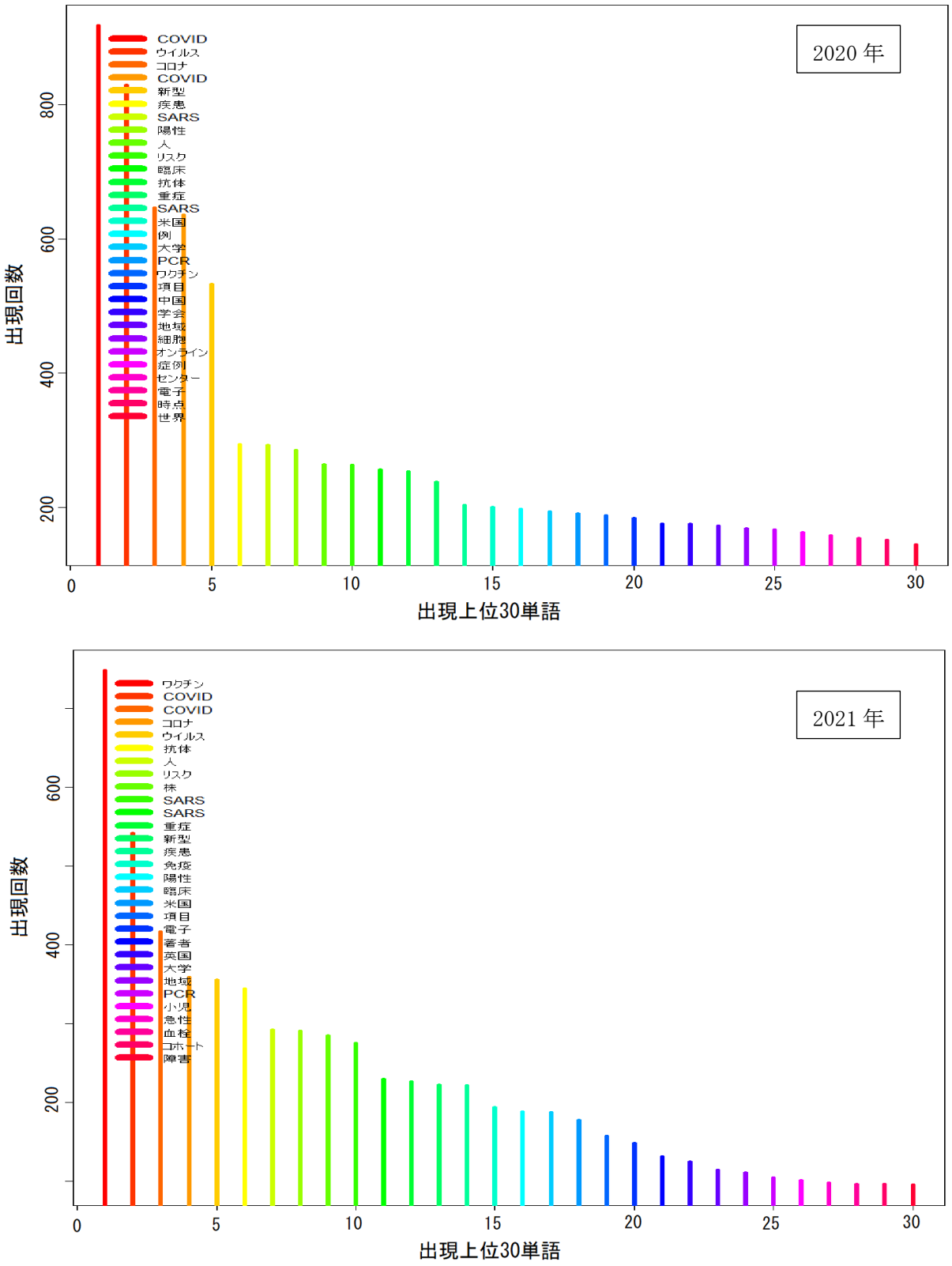



Fig. 1　テキストマイニングによる、1 単語における頻度解析結果（上：2020 年、下：2021 年）縦軸は、出現回数、横軸は出現上位 30 単語を示している。

2020 年


2021 年


Fig. 2　テキストマイニングによる、1 単語におけるワードクラウド
（上位 40 単語、上：2020 年、下：2021 年）
単語の出現頻度の高さに応じて、プロットされる単語のフォントサイズが大きくなる。単語はランダムに配置されている。

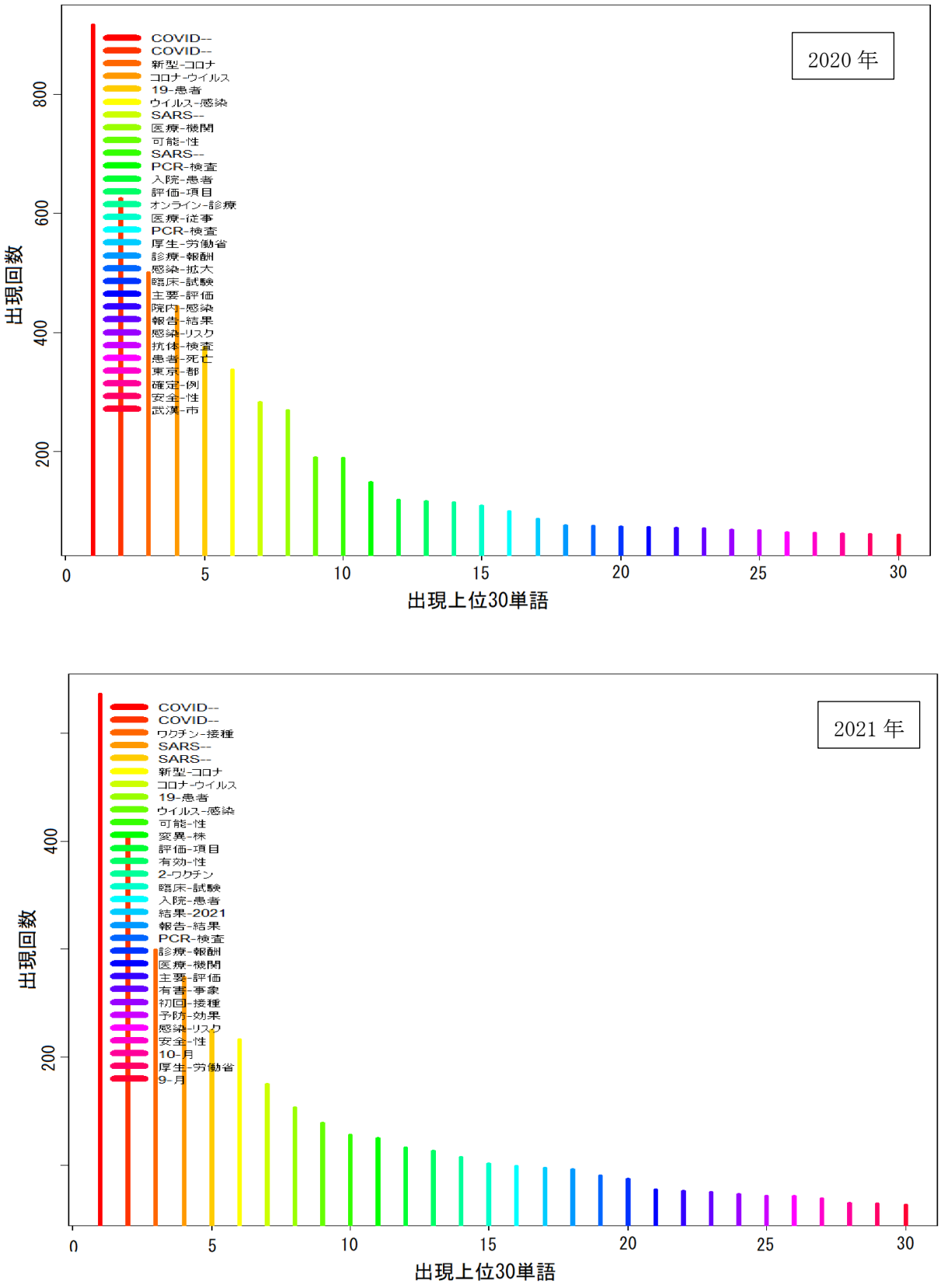



Fig. 3　テキストマイニングによる、２単語１節における頻度解析（上：2020 年、下：2021 年）縦軸は、出現回数、横軸は出現上位 30 単語を示している。

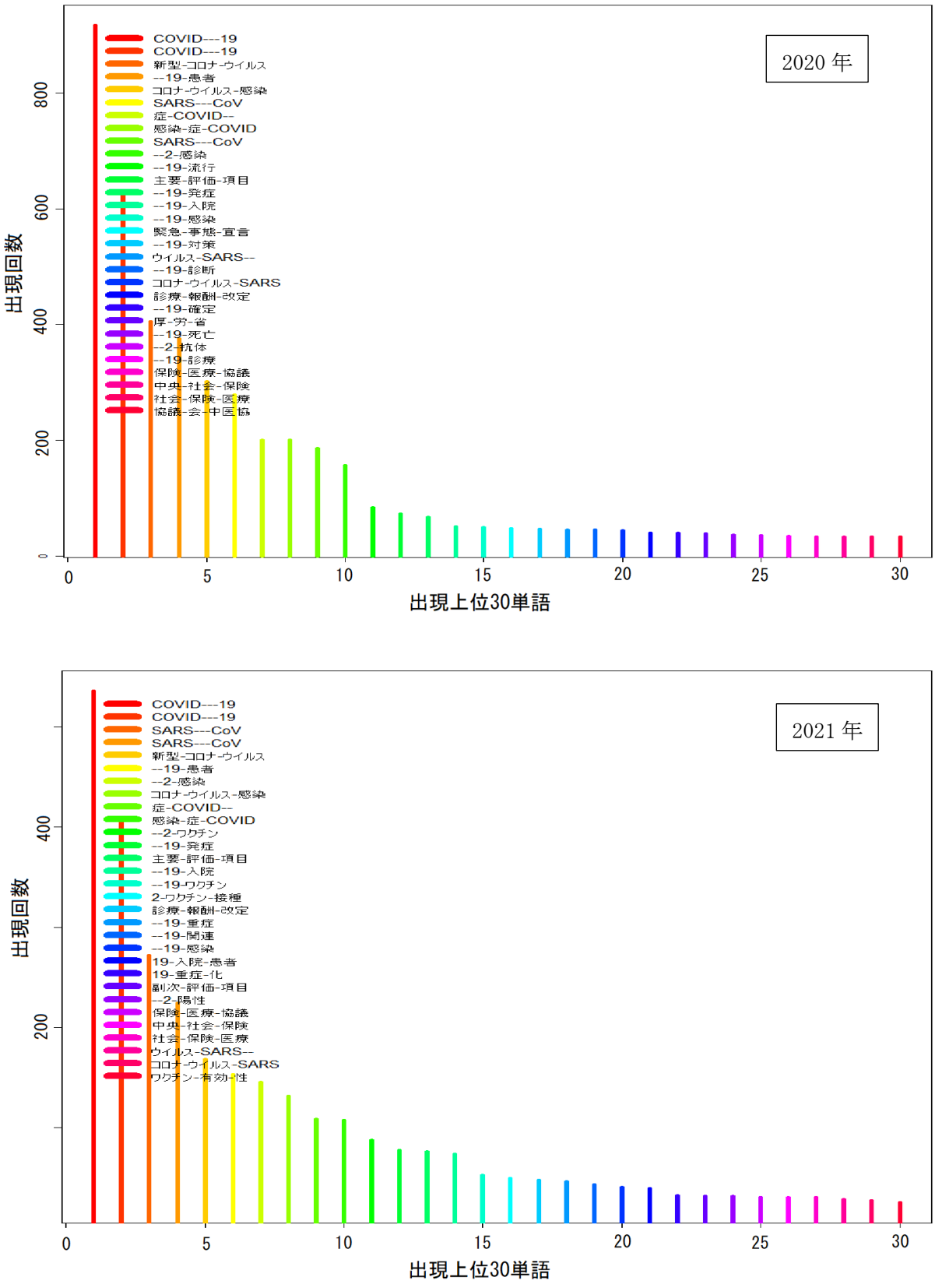



Fig. 4　テキストマイニングによる、３単語１節における頻度解析（上：2020 年、下：2021 年）縦軸は、出現回数、横軸は出現上位 30 単語を示している。

3. ネットワーク分析

テキストマイニングの分析方法の一つにネットワーク分析がある。これを行うことで、語と語のつながり関係や一つの文や段落における語の出現パターンの類似性や文字列間のリンクをもとに，文章中におけるそれらの語のつながり関係を共起ネットワーク図として可視化することが可能になる。また、各ネットワークの構成要素や関係性は全く別物であるが、これらのネットワークは複数の点で何らかの関係性でつながれていることも認識することが可能になる。そこで、今回は得られたテキストデータから特性の異なる２つの手法である Link Community（Fig. 5）および Overlapping Cluster Generator（Fig. 6）で解析し、それぞれ共起ネットワーク図を作成した。

Link Community からの共起ネットワーク図では、2020 年、2021 年ともに新型コロナウィルス感染症に伴うネットワークが多く形成されていることがわかる。2020 年および 2021 年のどちらにも“ウィルス”という単語がネットワークの中心にあるのでその部分に注目すると、2020 年では、“ウィルス”という単語をネットワークの中心に SARS、コロナ、感染などのワードがリンクしていることもわかり、さらにそこから次のワードに多くの共起関係がみられる（Fig. 5、上、赤線で囲った部分）。2021 年も“ウィルス”を中心に SARS、検出、RNA、感染などのワードがリンクしており、さらに 2020 年よりも多くのワードと共起され、より複雑なネットワークが次に形成されていることがわかる（Fig. 5、下）。1 年の経過により、このウィルスに関連する情報が多く報告され、Web 上に記載されたことがわかる。

Overlapping Cluster Generator からの共起ネットワーク図をみると、解析手法に違いはあるものの、例えばウィルスという単語に注目すると Link Community と同じようなネットワークが形成されていることがわかる（Fig. 6、赤線で囲った部分）。しかし、本手法を用いることで、より詳細なコミュニティーとネットワークも見ることができることがわかった（Fig. 6）。


Fig. 5 Link Community による共起ネットワーク図（上：2020 年、下：2021 年）
赤枠内は、“ウィルス”を中心としたネットワークを示している。

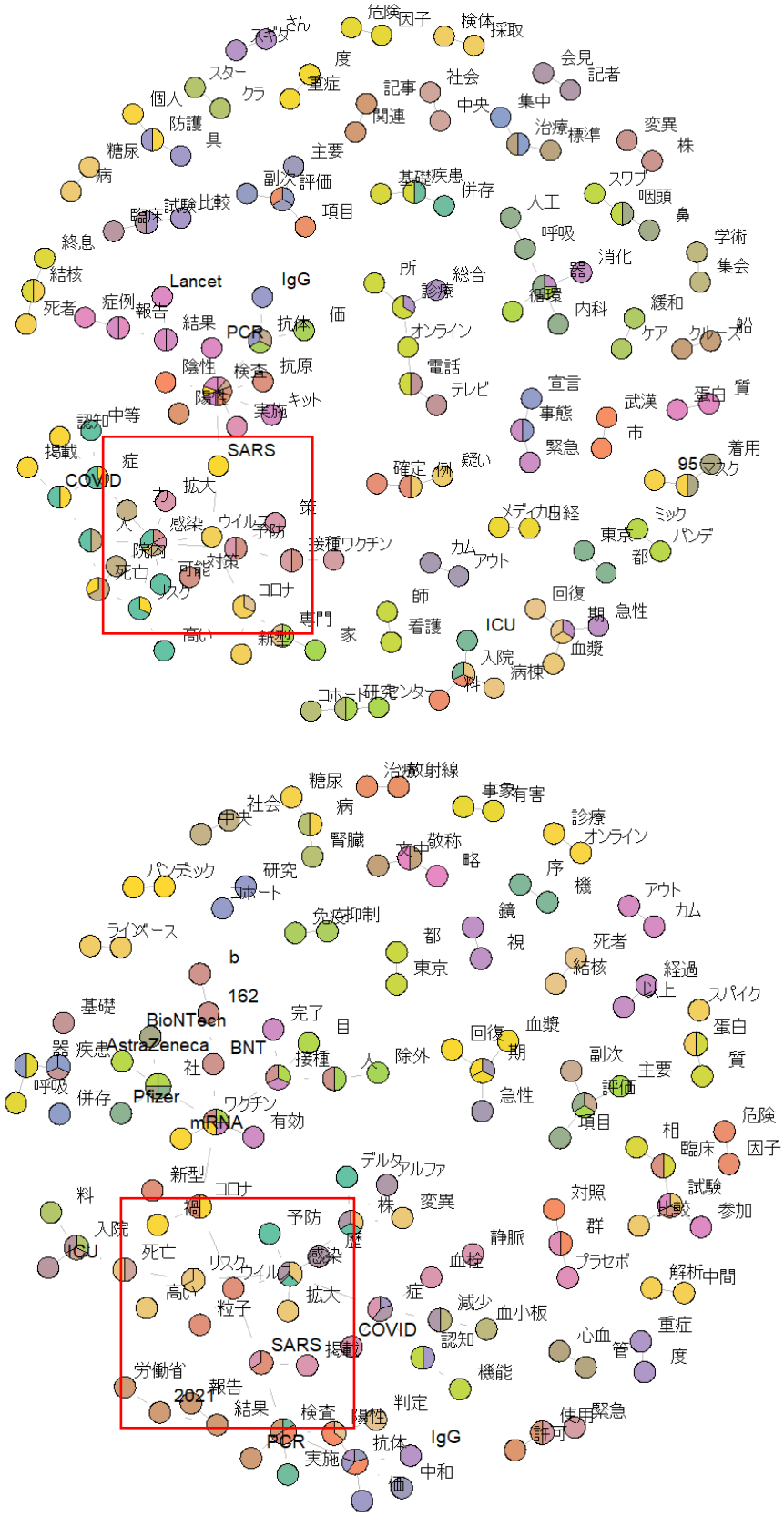



Fig. 6　Overlapping Cluster Generator による共起ネットワーク図
（上：2020 年、　下：2021 年）、
赤枠内は、“ウィルス”を中心としたネットワークを示している。

## 考察

本研究では、インターネットから医薬関連情報を得るため、日経メディカルの Web ページ中から 2020 年および 2021 年に記載された情報を Web スクレイピングによりテキスト抽出を行った。「newspaper3k」は、いずれも一般的なニュースサイトや SNS の記事を解析しており、今回のように、医薬関連情報収集に特化した解析を行っている事例は見当たらない。また、「newspaper3k」を用いたスクレイピングに関しては特にプログラムが動作しないといった不具合もなくテキスト情報を抽出することができたことから、今回のような条件下では十分活用できるライブラリであることが示唆された。今回は「newspaper3k」のみで検討したが、今後、他のライブラリについても比較検討したい。

また、今回は抽出したテキストデータをテキストマイニングにより頻度解析し、その年にどのような単語が多く使用されたかを判別した。当然、2020 年、2021 年ともに新型コロナウィルス感染症に関連する単語がほとんど上位を占める結果にはなったが、この結果から、医療系におけるそれぞれの年のトレンドとなるワードを解析し、視覚的なデータとして提供可能となることがわかった。医療系の単語は特徴的なものが多く、今回の結果でも「COVID」や「SARS」といった一部の単語について、同じ単語が別の単語としてカウントされている例が確認された。これに関しては解析に利用した「Mecab」の辞書機能を更新することでより精度の高い結果を得ることが可能になると考える。今後さらに検討していく必要がある。また、フィルターの設定を変化させ句読点やハイフンといった明らかに解析に不要な単語を除外することで、より自分が望む範囲のデータをピンポイントで入手することが可能になる。今後、辞書機能の更新やフィルターをかける単語を精査し、より医薬関連情報収集に特化したシステムを構築していく必要があると考える。

さらに本研究では、テキストマイニングからネットワーク解析まで展開した。今回抽出したテキストからは、新型コロナウィルス感染症に関連するネットワークが多くはなってしまったが、リンク数やコミュニティー数を十分にもった様々なネットワークの関連性を見ることができる共起ネットワーク図を示すことができた。使用した２つのネットワーク分析手法である「Link Community」と「Overlapping Cluster Generator」は、今回のような特定の Web サイトからの Web スクレイピングによって習得したテキストでもネットワークを可視化するために十分に適用可能であることを示すことができた。今回は「Link Community」と「Overlapping Cluster Generator」を選択したが、今後、特性の異なるほかの手法でも検討したい。

このようにインターネットから得られた膨大な情報量であったとしても、適切な解析をすることで、望むデータとして整理することが可能であることが示された。しかし、インターネット上のデータや情報は必ずしも正しいものとは限らない。情報の正確性に関してはデータ整理ののちに検証が必ず必要になる。検証の際は、信頼できる情報源からデータを取得しているか、古い情報を使用していないか、他のメディアでも同様に扱われているかといった点を確認しなければならない。今回、情報源として

使用した「日経メディカル Online」は、薬剤師会員数が 2020 年 3 月時点で 15 万人を超えているメディアであること、医療従事者全般に関する情報が掲載されているため幅広い領域から薬剤師が求める話題のトピックを抽出できるのではないかと考えたことから解析対象メディアとして採用した。しかしながら、医薬関連情報を掲載する Web サイトは他にも多数あり、1 メディアのみのデータでは結果に偏りがでる可能性も十分に考えられる。情報の正確性を担保するためにも複数メディアのデータ抽出も行い情報源の違いによるトレンド把握にどのような差がでるかを解析する必要がある。さらに、医薬品情報であれば、独立行政法人医薬品医療機器総合機構（PMDA；Pharmaceuticals and Medical Devices Agency）のホームページ[14]で公開されている情報と照らし合わせて検証し、入手した情報の正確性を確認することが必要になると考える。

今回の手法はこれからさらに改善していく必要性はあるが、薬剤師がインターネット上にある大量の医薬関連情報を簡便に入手するうえで十分に活用可能であると考える。また、今後は英文で記載された医学論文情報検索サイトからのテキストデータ抽出・解析も行い、世界レベルの最新情報についても簡便に入手できるかどうか検討していきたい。

## 引用文献


1) [internet] 総務省、「令和２年通信利用動向調査報告書」, https://www.soumu.go.jp/johotsusintokei/statistics/pdf/HR202000_001.pdf
2) Ryan Mitchell. Web Scraping with Python. O'Reilly Media, Inc.; 2015.
3) クジラ飛行机, 「シゴトがはかどる Python 自動処理の教科書」(第 3 刷) 株式会社マイナビ出版, (2021).
4) 斎藤 貴義. 「スクレイピング・ハッキング・ラボ Python で自動化する未来型生活」(Ver.1.1) 株式会社インプレス R&D, (2020).
5) Silge J, Robinson D., Text Mining with R. O'Reilly Media, Inc, (2017).
6) Hadley W, Grolemund G., R for Data Science, O'Reilly Media, Inc.; (2017).
7) 石田 基広, R によるテキストマイニング入門, (第 2 版) 森北出版株式会社, (2017).
8) Samatova N F, et al., Practical Graph Mining with R, CRC Press, (2014).
9) Caldarelli G, Shessa A., Data Science and Complex Networks, Oxford University Press, (2016).
10) [internet] Newspaper3k: Article scraping & curation, https://github.com/codelucas/newspaper
11) 井上 寛, テキストマイニングとネットワーク分析を用いたオープンキャンパスアンケート自由記述の分析, 第一薬科大学研究年報, 35, 45-53 (2018).(平成 30 年度 学内学術奨励金 成果報告)
12) Ahn Y-Y, et al. Link communities reveal multiscale complexity in networks, Nature; 466, 761-764 (2010).

13) Becker E, et al., Multifunctional proteins revealed by overlapping clustering in protein interaction network, Bioinformatics, 28, 84-90 (2012).
14) 独立行政法人医薬品医療機器総合機構 (PMDA; Pharmaceuticals and Medical Devices Agency) ホームページ, https://www.pmda.go.jp/index.html