Webニュースのアクセスランキングの時系列データから読み解く社会の変化

道越秀吾
（京都女子大学現代社会学部助教）
丸野由希
（京都女子大学現代社会学部准教授）

インターネットが普及した現在では紙媒体の新聞だけではなく、Webによって新聞記事を読むことも一般的になってきている。Webの新聞記事はアクセスランキングが発表されている場合もある。アクセスランキングから読者の関心を強く集める記事を観測できるため、社会的に注目された事象を抽出し社会情勢などを理解できる可能性がある。本研究の目的は、Web記事のアクセスランキングから得られる情報を可視化することによって、社会情勢などを容易に理解する方法を検討することである。2017年から2021年にかけてのWeb新聞記事の見出しについて、極性辞書による感情分析、ワードクラウドによる単語の使用頻度の表示などを行うことによって、解釈可能で意味が理解しやすい可視化を行うことができることがわかった。
キーワード：テキストマイニング、感情分析、ワードクラウド

1．はじめに

新聞記事は、国内外における社会動向や出来事を速報的に広く報じるものである。そのため、社会全体の流れや関心事などの情報がテキスト情報として含まれている。新聞はこれまでは紙媒体によって定期刊行される形態であった。しかし、インターネットが普及した現在、Webによって新聞記事を読むことも一般的になってきている。Web新聞記事の場合、スクレイピングによって機械的な情報収集が可能であり、それらに対してテキスト分析の手法を応用することによって多様な分析が可能である。そのため、これまで、様々な動機によりWeb新聞記事を対象とした研究が行われてきた。

Web新聞記事を対象とした研究例は非常に多く、ここで全て取り上げることはできないが、本研究と方向性が近く関連するものについて、いくつか取り上げる。まず、佐藤他（2009）は、災害に関するWebニュース記事を解析することで、災害の可視化を行うシステムを提案している。このシステムによって災害の社会現象としての側面が理解できるとしている。このような新聞記事を可視化することによって社会現象を読み解くというアプローチは広く一般的に応用できる。

宮原（2019）は、アクセスランキングを活用して、地域ニュースに特化した分析を行っている。アクセスランキングという概念は、紙媒体の新聞しかなかった時代では存在しえなかったものである。従来の紙媒体において記事の重要性は、一面の見出しとして扱われるかなどで判断できただろう。しかし、一面の見出しとするかどうかについては、発信側、すなわち報道機関が決めている。一方で、アクセス数ランキングは読者側から見た記事の注目度・重要度を得ることができる。Webで記事が読まれることが一般的になった時代では、アクセス数ランキングこそがニュースの価値を決めていると考えることができるかもしれない。

そこで、本研究では、ニュースサイトのアクセス数ランキングを用いて、社会現象を読み解くための可視化手法を検討する。2017年から2021年ま

での５年間を対象に絞り分析を行った。この５年間の特徴的な社会現象というと、新型コロナウィル感染症（COVID-19）の広がりや、東京オリンピックなどがあった。これらが Web 新聞記事の中でどのように扱われたかをテキスト情報から読み解き可視化することによって、社会全体の流れを分析し理解する手法を確立することがこの論文の目的である。

本研究では、ニュース記事の内容のエッセンスが集約されていると考えられる見出しを分析対象とする。今回は、対応する記事の本文については分析対象としなかった。見出しにどのような単語が多く用いられているかを分析することによって、その時の社会情勢を要約する。そして、本研究では極性辞書を用いた感情分析も併せて行なった。ニュース記事を対象とする感情分析の先行研究はいくつかある（e.g., 熊本他 2005，高野他 2011，高津他 2021）。本研究では、極性辞書によって、多くアクセスされた新聞記事がポジティブであるかネガティブであるかを定量的に求め、その時の世相を表すことを試みる。

以下、この論文の流れである。２章では、用いたデータセットについて説明する。３章では、解析手法の概要を説明する。４章、５章は感情分析の結果、見出しの中の単語の頻度分布の解析結果を述べる。６章は、本研究で得られた結果をまとめ、今度の展望を議論する。

２．データセット

本研究では、2017年９月21日から2021年７月20日の朝日新聞デジタルの Web サイト「朝日新聞デジタル（http://www.asahi.com/whatsnew/ranking/）」のアクセスランキングデータを用いた。朝日新聞デジタルのアクセスランキングでは、15分ごとに更新されるリアルタイムランキングと１日前のアクセスランキングであるデイリーランキングの２種類がある。ともにアクセスランキング上位10位までが掲載されている。今回の解析ではデイリーランキングを使用した。

Web サイトに掲載されているデイリーランキングの見出しデータをスクレイピングを用いてデータ取得を行った。プログラミング言語 Ruby とそのライブラリである Nokogiri を用いてアクセスランキングの順位・タイトル・URL を取得した。Nokogiri は Web スクレイピングで使われる Ruby のライブラリで、CSS セレクタを用いて、HTML の構造を解析し、特定の要素を抽出する。毎日23時にスクリプトを実行するためのデーモンプロセスである cron を用いてデータ取得のためのスクリプトを実行し、CSV フォーマットで保存した。これにより、長期間のアクセスランキングの変動の解析が可能となった。

３．手法

3.1. 形態素解析

形態素解析とは、文を単語（≒形態素）単位に分割し、各単語の品詞を特定する解析技術のことを指す。例として、「庭には二羽ニワトリがいる」という文章を形態素解析すると以下のようになる。

庭（名詞）／に（助詞）／は（助詞）／二（名詞）／羽（名詞）／ニワトリ（名詞）／が（助詞）／いる（動詞）

本研究では、オープンソースの形態素解析エンジンである MeCab を用いた。MeCab は、形態素解析エンジンの中でも、高速かつ高精度の解析を実現している（Kudo et al. 2004）。新聞記事という特性上、新語や固有表現が多く登場する。しかし、MeCab で標準的に用いられる IPA 辞書では新語や固有表現への対応が不十分であるため新語の反映が早いという特徴のある mecab-ipadic-NEologd を辞書に採用した。この辞書では、Web の言語資源から得た新語を随時追加することによって、数多くの新語や固有表現に対応している（Sato 2014，佐藤他 2015，2016）。

3.2. ワードクラウド

ワードクラウドは文章中で出現頻度が高い単語を複数選び出し、その頻度に応じた大きさで図示する手法である。出現頻度が多いほど、フォントサイズが大きくなる。単語の使用頻度可視化によって概況が把握しやすくなる特徴があるため、文章の要約などに用いられることがある。本研究では、プログラミング言語 Python のライブラリを用いた。このライブラリでは、文字の大きさは、

出現頻度と出現頻度の順位の両方を考慮して、サイズが計算される。

3.3. 極性値辞書

大量のテキストから感情を機械的に読み解くには、書かれた内容の感情を評価するための辞書が必要である。感情表現の細かな分類を行なった研究事例もあるが、本研究では感情を2種類のみに分類する極性辞書を用いる。極性辞書では、与えられたキーワードごとに、ポジティブ（望ましい）かネガティブ（望ましくない）であるかを表す極性の情報が付随する。

約5000件の評価表現リストを持つ極性辞書を用いた（小林他 2005，東山他 2008）。MeCabによって得られた単語に対して、極性辞書を参照し、極性の属性を得る。見出しの中で、極性が定義される単語総数tとして、ポジティブな極性を持つ単語数をp、ネガティブな極性を持つ単語数をnとする。極性が定義される単語総数tに対するpとnの差の割合によって、与えられた見出しに対する極性値を定義する。つまり、

$$(p\text{-}n)/t$$

が見出しの極性値である。定義によりこの値は－1以上1以下の範囲である。もし、辞書に該当する単語が存在しない場合は0と定義する。0の時がニュートラル、または判別不可能であり、ネガティブ、ポジティブの極性は、極性値の符号と対応している。

1日あたり10件の見出しがあるため、各々の見出しの極性値を計算して、それらの平均値を指定の日の平均極性値として定義される。これも同様に－1以上1以下である。これをデータセットが保持されている期間の全ての日に対して計算を行うことによって、極性値の時系列データが得られる。

4．極性値による感情分析

2017年9月21日から2021年7月20日までの日々の極性値のヒストグラムが図1である。日ごとの極性値の値はばらつきがあり、最小値－0.76から最大値0.57までの範囲で分布している。平均値は－0.2179，中央値は－0.2343、標準偏差は0.1934である。グラフより0を含むビンの頻度が、それに隣接するその左右のビンよりも高くなっていることがヒストグラムより読み取れる。極性値が0になるケースとして、見出しに含まれる単語が1つも辞書に該当しなかったというケースがある。1日あたり10本の見出しの全てが0ではない極性値をもたなかったことが多かったということを意味している。

また、平均値－0.2179であり、全体的な傾向として、極性値としてはネガティブに偏っていることがわかる。この理由を確かめるために、上位300位までの使用頻度の高い単語について、ポジティブ、ネガティブそれぞれに該当する単語に分けて、それぞれ上位から順に示した。結果は以下の通りである。

（ポジティブ）
結婚，人気，命，ため，長，俳優，大統領，可能性，話題，笑顔，人生，救助，支援，感謝，優勝，希望，幸せ，金，保護，若者，奇跡，絵，記録，情報，ファン，名，勝利，成功
（ネガティブ）
死亡，感染，逮捕，容疑，批判，死去，疑い，被害，死，事故，虐待，自殺，台風，中止，殺害，事件，遺体，問題，衝撃，病院，性暴力，感染者，新型肺炎，抗議，けが，注意，原因，火災，死者，患者，処分，がん，衝突，閉店，指摘，悲鳴，離婚，暴行，転落，不正，反対，地獄，過去，ウイルス，炎上，恐れ，ミス，懸念，反発，停電，拒否，閉鎖，傷，恐怖，地震，混乱，セクハラ，違和感，告発，不安，怒り，限界，いじめ，炎，嫌，放置，異変，暴走

なお、極性辞書に該当しなかった語はここでは示していない。ここで示した単語リストのうちで、ポジティブの単語グループは19個、それに対してネガティブは39個であり、ネガティブな単語の方が2倍近く多く用いられている。ネガティブは、死亡、逮捕、被害などといった事件報道でよく用いられる単語が上位に位置している。そのため、全体として、ネガティブな単語に偏った語彙が用いられているようである。ニュースは社会的な影

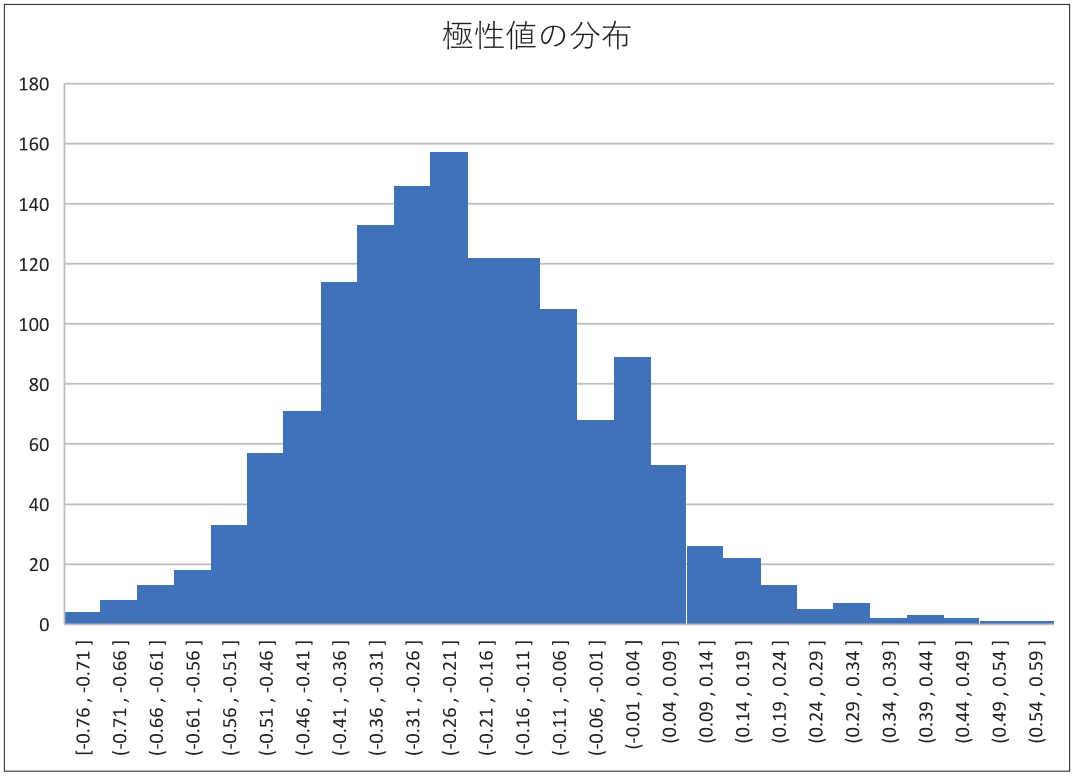

図１　極性値のヒストグラム。2017年９月21日から2021年７月20日までの日毎の極性値の頻度分布である。

響の大きな事件や事故などを扱うことが多い。そのため、一般的な文章と比べてネガティブと見なされる語彙が多く扱われるということが一因として考えられるだろう。

図２は、日ごとの極性値の時間変化をプロットしたものである。日々の極性値だけでは分散が大きくトレンドが掴みにくいため、移動平均も合わせてプロットした。移動平均を計算する区間の長さは30日、つまり、およそ１ヶ月の期間である。

この図より2020年１月から2020年９月にかけて、全期間平均に比べて移動平均値が低い値となっていることが読み取れる。つまり、このおよそ８ヶ月の期間はネガティブな単語が多く用いられた期間である。この期間で特に多く使われた単語のリストを抽出した。

（ポジティブ）
ため，俳優，話題，長，旅行，可能性
（ネガティブ）
感染，新型肺炎，感染者，死亡，逮捕，批判，死去，容疑，病院，死，ウイルス，殺害，患者，中止，死者，虐待，閉店，悲鳴，自殺，嫌，疑い，自粛，解雇，注意，危機，恐怖，怒り，不安，炎上，事件，被害，抗議，衝撃，処分

この期間における使用単語の上位200位のうちで、ネガティブまたはポジティブの属性が付与されているものを抽出し上位から並べたものである。ポジティブな単語の数はわずか６個であり、一方でネガティブの単語数は34個であった。期間全体ではポジティブな単語数に対してネガティブな単語数が約2.05倍であったが、この2020年の期間に絞るとポジティブな単語数に対するネガティブな単語数の割合は約5.67倍と偏っていることがわかる。

そして、その単語の中身を見ると、感染、新型肺炎、感染者、病院、ウィルス、患者といった、新型コロナウィルス感染症に関連した言葉が上位

にきていることがわかる。その他は通常の事件報道で用いられるような死亡、逮捕、批判、死去、殺害などである。

全期間の平均がネガティブに偏っている理由は、新聞記事であるという特性上、事件や事故の報道に関するキーワードが多いため全体としてはネガティブに偏っているのではないかと議論した。2020年の期間を対象に絞った場合は、事件・事故の報道に加えて、コロナウィルス感染症というネガティブで社会的影響の大きい事象が繰り返し報道されたためこのような結果になったのであろうと考えられる。

全体的な傾向の中でも一時的に極性値が増加に転じることもある。例えば、直近の2021年１月から３月は全体平均より高い数値になっている。この期間において使用された単語のリストは以下の通りである。

（ポジティブ）
人生，大統領，長，認定，救助，可能性，生活保護，若者，根拠
（ネガティブ）
感染，死亡，逮捕，批判，問題，違反，性暴力，津波，死去，被害，餓死，疑い，事故，震災，理不尽，容疑，がん，中止，重症，処分，死，リコール，中断，病院，疑惑，死者，弾圧，不正

やはり、ネガティブな単語の方が相対的に多く用いられている状況は変わりない。この時期においてポジティブな単語として最も多く用いられて

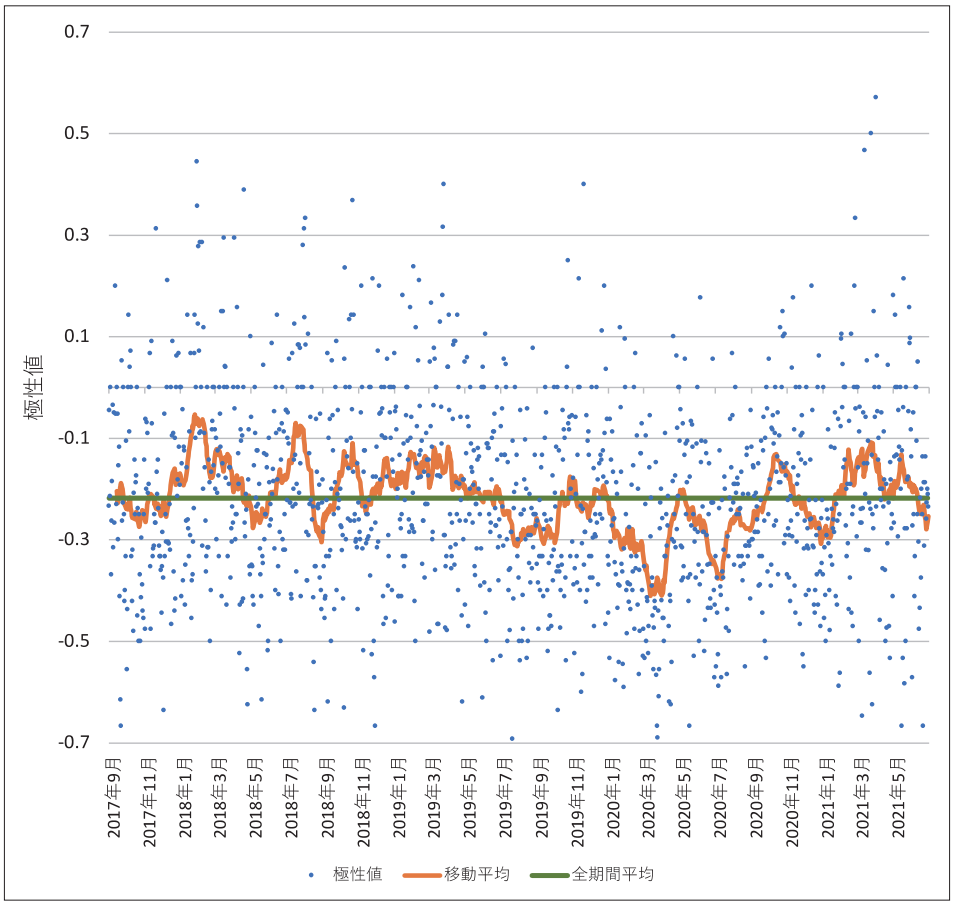

図２　日々の平均極性値の時間変化。青い点が日々の極性値である。橙色の実線が、期間30日の移動平均、緑色の実線が全期間を通した平均値である。

いる人生という単語は、アクセスが多かったコラム記事の中のタイトルとして使われた単語のようである。また、大統領については、ちょうどこの時期にアメリカ合衆国の大統領選の結果が出た時期であるため盛んに報道された。本研究で定義する極性値は相対的な意味であるため、極性値が平均より高かったという場合は、ポジティブが増えた、ネガティブが減ったという２つの要因に分けて考えることができる。大統領選やコラムへの注目度の高まりといったコロナウィルス以外のポジティブな事象が発生したということ、コロナウィルス感染症の問題は長期化していたものの、時間が経ち慣れなどから関心としては薄れる傾向にあった、などが要因として考えられるだろう。

ここまでの議論で、新聞記事の見出しに対する極性値を計算することで、その時の世相や情勢を定量的に１つの数値として得られた。そして、それが概ね実際の事象と対応させて解釈することができそうである。しかし、極性値だけでは、その極性値となった要因がどのような出来事であったか、なぜそのような数値になるのかを知ることはできない。そのため、次の章で、見出しに登場する単語の頻度を解析し可視化することで、その時の社会情勢の概要を掴むことができるか検討する。

５．単語の出現頻度

5. 1. 単語の出現頻度の分布

まずは、基礎的な解析として、見出しに登場したすべての単語に関する頻度分布を議論する。見出しの中で特によく使われた単語の概要を掴むために、全ての見出し中における単語の登場回数をカウントし、上位100位までの単語を抽出し、使用頻度順に並べた。結果は以下の通りである。

さん，人，氏，コロナ，死亡，感染，女性，歳，年，円，逮捕，日本，中，ら，万，母，男，首相，私，後，容疑，父，写真，車，男性，妻，マスク，娘，の，前，日，千，五輪，詳報，容疑者，性，韓国，批判，発言，死去，夫，会見，米，疑い，涙，新型，被害，動画，東京，選手，者，息子，まとめ，こと，子，死，今，事故，中国，時間，発見，初，目，手，発表，何，公開，社長，声，理由，虐待，宣言，結婚，店，億，陛下，世界，自殺，思い，先，トランプ氏，知事，職員，心，彼女，医師，分，台風，謝罪，記者，暴力，女児，時，少女，ぶり，数，国，新た，客，確認

最上位の３位以内は、さん、人、氏などで、敬称や人数を表す単位であり、新聞記事では一般的に用いられる言葉である。次に４位に位置するのがコロナである。2017年以降の見出しの解析であるが、2020年から広がりを見せたコロナが、一般的な言葉を除けば最上位にきており、いかに社会的な関心を引いた事象であるかが伺える。さらに６位には感染など、コロナウィルス感染症関連の言葉が見られる。これらの単語に関連する詳細な分析は５．２節以降で論じる。

続いて、全期間に関する出現する単語の頻度の分布の性質を調べた。図３は、対数順位に対する対数頻度の関係を示している。対数は自然対数で計算した。すると、両対数グラフにおいて、対数順位が２から6. 5の範囲で概ね直線で表されることがわかる。最小二乗法によって、適合する直線を求めた。順位を r、頻度を f とすると、

$$\log f = -0.643913 \log r + 7.34224$$

と表される。傾き及び切片の p 値は 2×10^{-16} 以下であり、調整決定係数は0. 9963である。式変形をすると、r, f の関係はべき則に従うことがわかる。つまり、

$$f = 1544.16\, r^{-0.643913},$$

の関係が成り立つ。

一般に文章の中のある単語の出現頻度を考えて、ある単語の出現頻度が上位から r 番目である時、その出現頻度と r の積が一定であることが知られており、ジップの法則と呼ばれている（Zipf 1935）。これは、出現頻度が $1/r$ に比例することを意味している。今回フィッティングによって得られた結果はべき則であり、その指数としては反比例を意味する－１よりやや浅い数値となっているが、ジップの法則を一般化・拡張したものの一種であるとみなすことができる。以上は、比較的上位の領域である、対数順位が6. 5以下の場合であるが、対数順位が6. 5より大きい場合は、べき指数より急になっていることも読み取れる。

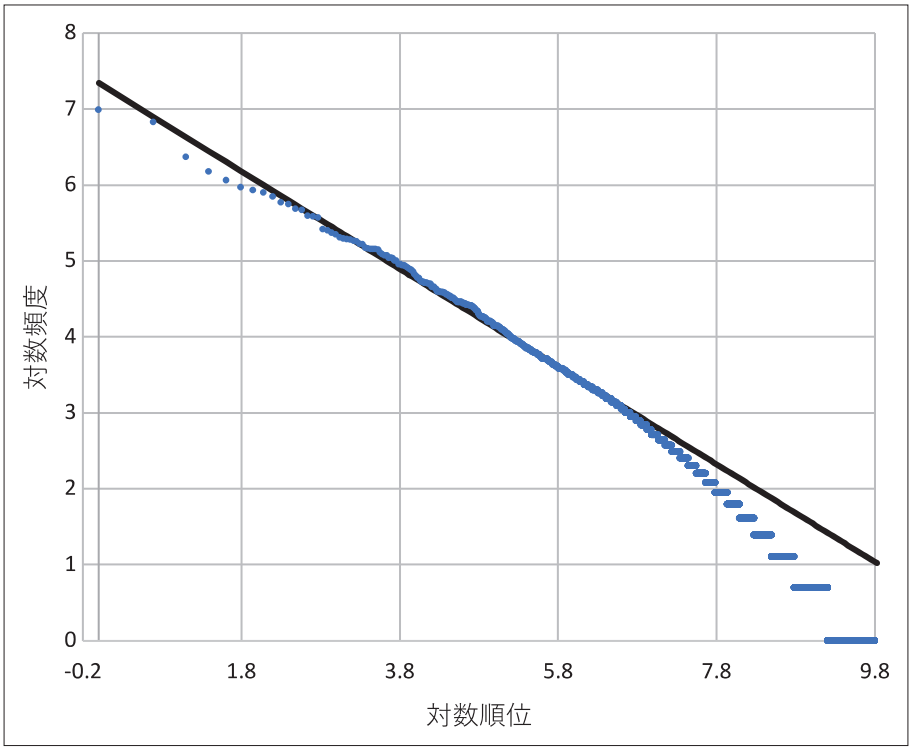

図３　対数順位と対数頻法則の関係。除去単語の頻度の順位の自然対数と頻度の自然対数の関係を青い点でプロットした。なお、実線はフィティングによって求めた直線である。

5. 2.　ワードクラウドの解析

図４から図７は、2019年度下期（10月から翌年３月）から2021年度上期（４月から９月）の半期ごとのアクセスランキングのワードクラウドである。コロナ関連の単語に着目すると、2019年度下期（図４）では「新型肺炎」「感染」「マスク」といった単語が出現している。2020年度上期から2020年度下期（図５、図６）では、「コロナ」「感染」「マスク」といった単語が出現しており、数あるニュース記事の中でもコロナ関連の記事への関心が高かったことを示唆している。また、2021年度上期（図７）では、「ワクチン」「接種」いう単語が見られるようになり、ワクチンに関する記事の出現およびアクセスランキング上位にあがってきたことを示している。

5. 3.　単語の出頻度の変化

１日あたり10本の見出しがあるが、それらの見出しに指定の単語が含まれる個数を数えて、各日毎の出現頻度を算出した。図８から図13は、コロナ、感染、新型肺炎、マスク、ワクチン、五輪などの言葉が使われる頻度の時間変化をそれぞれプロットしたものである。トレンドを把握しやすくするために、移動平均も合わせてプロットしてある。移動平均の区間の幅は10日である。

まず、図８はコロナという単語についてである。2020年１月以前には、見出しでコロナという単語の使用は見られなかったため、2020年１月以降の表示に限定した。日本においては、コロナウィルス感染症は2020年１月に初の感染者が確認された。その後の感染の広がりと共に報道が加熱していったことが見出しの出現頻度からもうかがえる。

2020年１月の段階で、すでに感染の広がり懸念され社会的注目が高まっていたと考えられる。例えばクルーズ船におけるコロナウィルス感染症の集団感染が発生したのは2020年１月末でありこの時期にはすでに社会問題となっていた（e.g., 国立感染症研究所「現場からの概況：ダイアモンドプ

図４　2019年度下期（10月～翌年３月）のアクセスランキングのワードクラウド

図５　2020年度上期（４月～９月）のアクセスランキングのワードクラウド

リンセス号における COVID-19症例」)。しかし、コロナという単語自体の使用数については１月では０であり全く使われておらず、２月から徐々に使われ始めている。この点を確かめるために、関連語である感染、そして新型肺炎という言葉で同様のプロットをしたのが図９と図10である。まず感染という言葉は2020年１月25日から登場している。コロナよりも前から登場しているため、感染症の報道自体は１月から確かにあったようである。しかし、その名称としてコロナという単語が用いられてはいなかった。次に新型肺炎という単語が初めて登場したのが2020年１月23日であり、１月31日では10本の見出しの中で４回登場しており、盛んに報道されていたことがわかる。ところが、2020年２月15日をピークに急速に登場頻度が低下し、2020年３月16日以降は登場していない。つまり、新型肺炎と入れ替わりのようにコロナという単語が使われるようになったようである。

図 6　2020年度下期（10月～翌年 3 月）のアクセスランキングのワードクラウド

図 7　2021年度上期（ 4 月～ 9 月）のアクセスランキングのワードクラウド

2020年 1 月の段階では、このコロナウィルス感染症で深刻な肺炎を引き起こすことが特にクローズアップされて報道されていた。しかし、病態が明らかになるにつれて、感染しても肺炎を引き起こさず軽症で済むケースも多く見られることがわかってきたことから、新型肺炎という呼称は実態と合わないということが認識されてきたためと考えられる。原因ウィルスを名称に含めて幅広く意味をとることができる方がより適切であることから、新型コロナウィルス感染症、つまりコロナという呼称に変わっていたのだろう。

コロナの出現頻度は 2 月以後急速に増加し、 4 月 6 日に移動平均の最大値となっている。この日だけで見出しの中に 5 回登場している。この日の10件の見出しは以下の通りである。

スマホ越しの「さよなら」　コロナ感染の母に娘は告げた

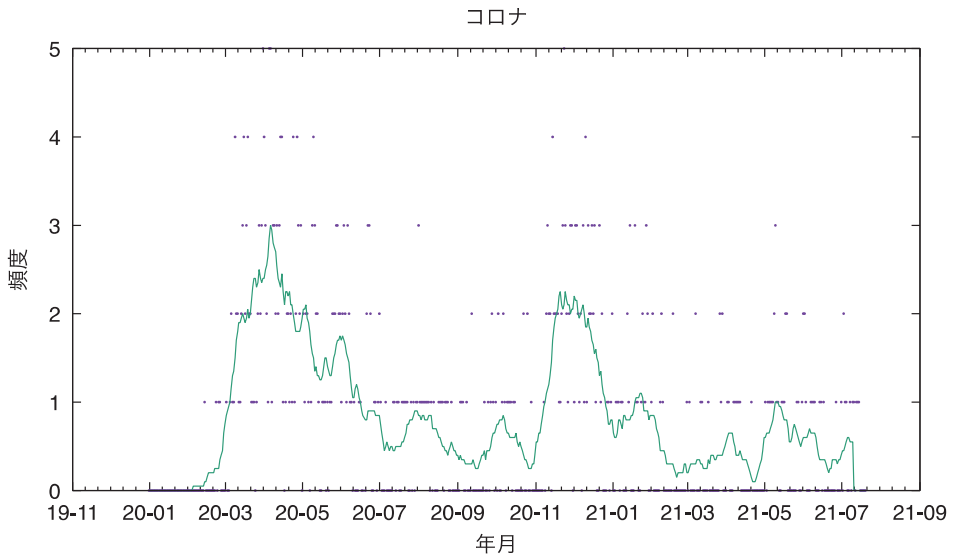

図 8　見出しにおける単語の出現頻度の時系列変化。点で示したものは各日の10本の見出しにコロナという単語の総登場数をプロットしたものである。また、緑色の実線は、幅10日に移動平均を示している。

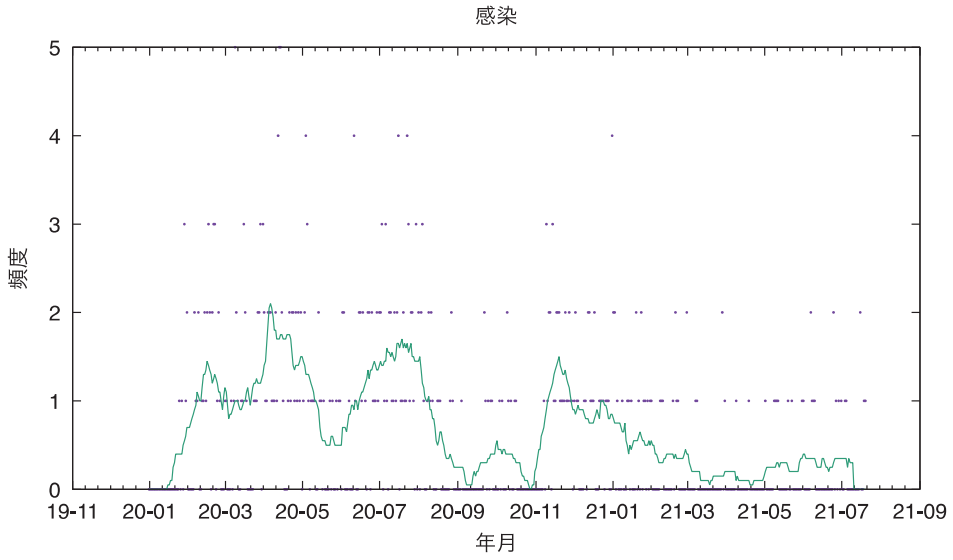

図 9　図 8 と同じであるが、感染という言葉についてプロットしたものである。

緊急事態宣言で何が起きる？　ロックダウンとの違いは
埼玉）新型コロナ　経路不明多く　東京近郊に多い傾向
父の人工呼吸器、電話で「若者に回す」　命の選別に絶望
緊急事態宣言の「時期は近い」　政府内で高まる容認論
空気感染しないコロナ、換気なぜ必要？漂う粒子を見ると
坂本龍一に清志郎が警告していた　コロナ危機「その後」
東京の感染倍増する時間「欧州に近い」　専門家の見方

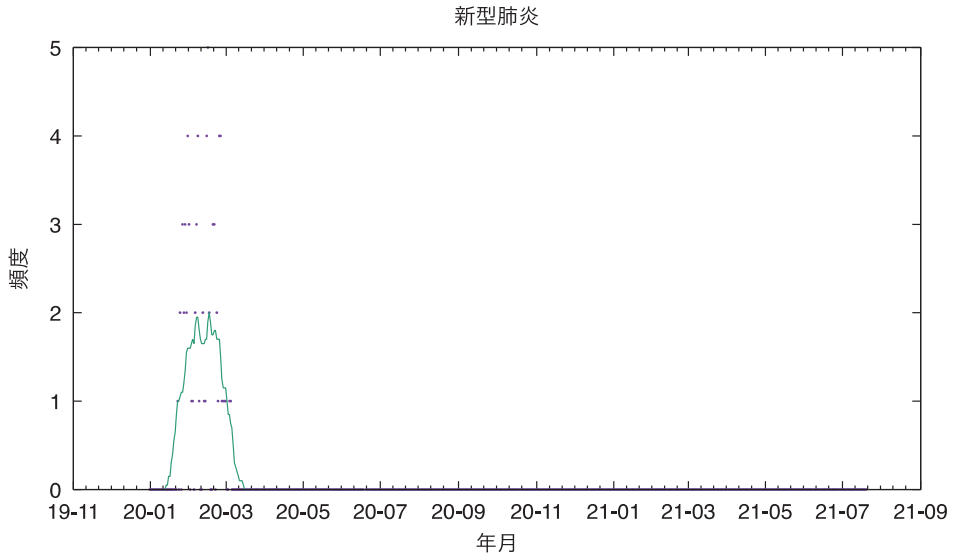

図10　図 8 と同じであるが、新型肺炎という言葉についてプロットしたものである。

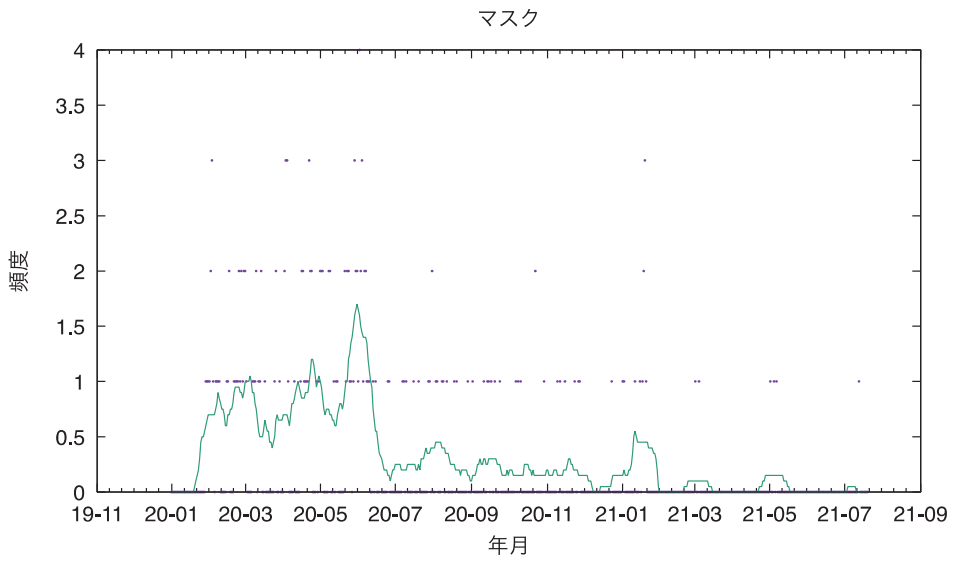

図11　図 8 と同じであるが、マスクという言葉についてプロットしたものである。

コロナ対応した北海道教育長が急死　62歳、循環器不全
持病ない10代が突然…欧州で相次ぐ死亡例、WHO 警鐘
引用元：asahi.com（朝日新聞 DIGITAL）の2020年 4 月 6 日のアクセスランキングの見出し

具体的にコロナという単語が含まれていたのは10件中 5 件の見出しであったが、コロナが含まれていない他の 5 件もコロナウィルス感染症関連の見出しであり、この時コロナウィルスに対する関心がピークを迎えていたと考えらえるだろう。

その後、見出しの登場数は、一時的な減少や増加を繰り返すものの移動平均が 0 となることはなく数日に 1 回はコロナという単語を含む記事へのアクセスが多く行われていたことがわかる。

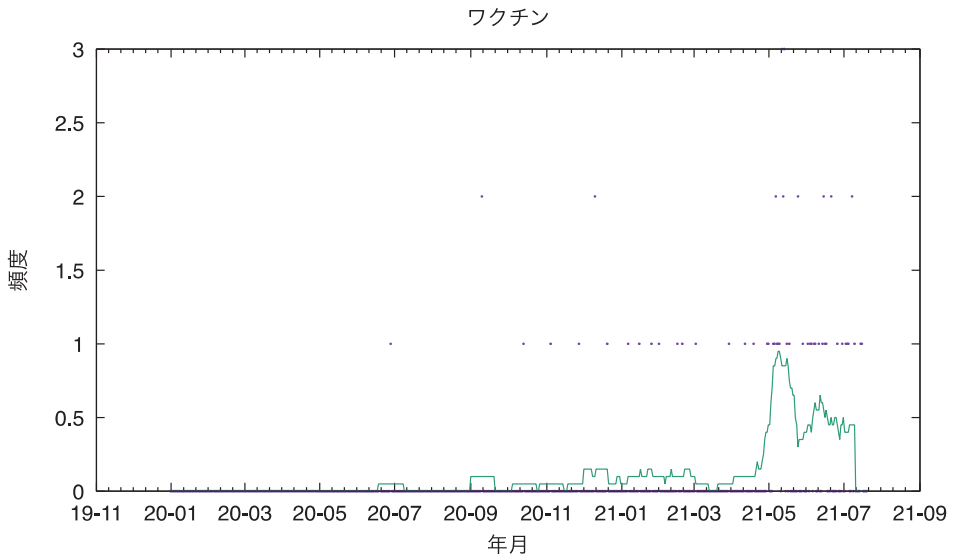

図12　図 8 と同じであるが、ワクチンという言葉についてプロットしたものである。

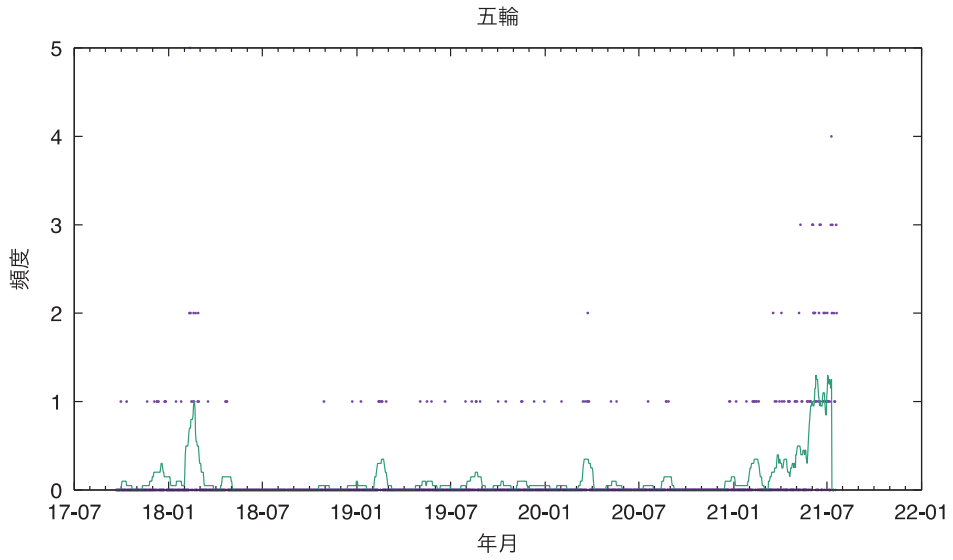

図13　図 8 と同じであるが、五輪という言葉についてプロットしたものである。

2020年の年末に 2 回目の出現頻度のピークが見られる。 2 回目の極大値となったのは、2020年11月25日であり、移動平均の値は2. 25である。つまりこのあたりでは 1 日平均して約 2 回はコロナという単語が含まれる記事があったことを意味する。2020年11月25日の周辺の見出しの実際の内容を確認したが、特にコロナに関連して集中して報道される重大な事件や出来事は確認されなかった。コロナに関連した相異なる事象が単発的に報道されていたに過ぎない。ではこの時期になぜコロナに関する記事がアクセスランキングの上位に来たのかを検討するため、新規感染者数との関連を調べた。図14は日本全体の新規感染者数の時間変化を表すグラフである。データは厚生労働省の提供するオープンデータを利用した（厚生労働省 新型コロナウィルスについて）。2020年11月末は、感

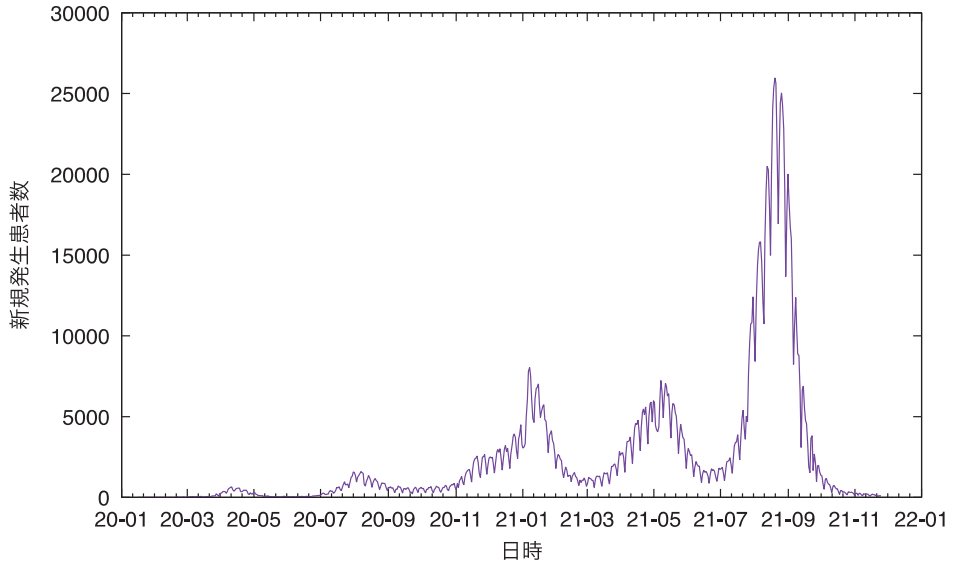

図14　新型コロナウィルス感染症の日毎の日本全体の新規発生患者数。

染の再拡大の兆候が見えてきたところであり、ピークは2021年１月８日であった。感染拡大への懸念が再び見えてきたところで、読者のコロナウィルス感染症への関心が高まりアクセス数としても高まった可能性がある。

新規感染者数の推移としては2021年５月にも再拡大が見られる。この辺りの時期では、コロナが含まれる記事の見出しは確かに増えるもののそれ以前に比べて最大値は低くなってきている。

コロナという語に関連して登場数の上位のキーワードとして、マスクとワクチンがあった。マスクの結果は、図11に示した。マスクという語は感染の拡大が始まった2020年１月から直ちに記事の見出しとして現れていたことがわかる。内容を確認すると、感染予防に対するマスクの効果の専門家の意見を掲載する記事やマスクの品薄に関連する報道が行われていた。感染急拡大の中、マスクが品薄となる副次的な社会問題が発生し世間の関心が高まっていたがわかる。この傾向は2020年６月ごろまで続き７月に入ると急速に減少している。５月から６月は、様々な企業がマスクの生産や販売に乗り出したことや、国によるマスクの配布などがあり注目が集まったようである。しかし７月にはマスクの品薄が解消されたため新聞記事としても報道されることが少なくなったようである。

ワクチンについての結果は図12である。感染流行の初期段階、つまり2020年の段階では、数としては少なく、１ヶ月に数本程度の記事で報道されるだけであった。それらの内容はワクチンの開発状況の報道である。しかし、2021年４月に入ると急速にワクチンに関する報道が多くなっている。国による希望者へのワクチンの接種が本格的に始まった時期に対応しており、接種予約や副反応、効果などワクチンに付随する様々な事象が報道されていた。それらによって、社会的関心が高まっていたことがわかる。

コロナ以外でこの調査期間を代表する事象としては、オリンピックがある。五輪の結果は図13である。オリンピックは2020年開催の予定であったが2021年に延期となった。延期決定はコロナウィルス感染症が広がりを見せた３月であった。延期そのものは多くの関心を引かなかったようであり、確かにピークを見せているものの、その最大値は大きくない。実際に開催された2021年７月には、多くの報道がなされていることがわかる。

ここまでの結果より、新聞報道の見出しのテキストデータにより、その時の社会情勢や関心の移り変わりを可視化し読み解くができることがわかる。

6．議論とまとめ

この論文では、Web 新聞の記事のテキストデータを解析し可視化することで、社会情勢を読み解き理解できるか検討した。

Web 新聞記事に関連する情報の中でも、アクセス数ランキングに着目した。従来の紙媒体の新聞記事では得ることのできない情報であるアクセス数ランキングは、記事の重要度・注目度を客観的に測ることができる指標であると捉えることができる。今回の解析では、内容が集約されていると考えられる見出しに注目し解析を行った。

まずは、社会情勢全体を定量化するための感情分析を行った。感情分析とは使用される単語から、どのような感情であるか分類する手法であるが、今回はポジティブとネガティブという２つに分ける極性の分析を行った。全期間を通した極性値の平均値は、０を下回り負の数となった。つまり、全体的にネガティブに偏った語彙が多く使用されていたということを意味する。新聞記事であるという特性上、事件や事故の報道が多く使用される語彙もネガティブに偏っていたと考えられる。しかし、その一方で、時期によって平均値からの短期的なずれが見られる。ネガティブよりもポジティブに偏る時期、逆によりネガティブの方が多い時期などがある。これらの情報より、見出しの極性値から、全体としてネガティブであるかポジティブであるかの大まかな傾向を掴むための有力な指標であるといえよう。

極性値の分析は容易に期間を通しての傾向や変化を掴むことができるため、全体的な解析や変化を理解するには、有力な指標であると考えられるが、その時に実際に何が起きていたのか、どう社会情勢が移り変わっていたのか、その中身について理解することができない。実際の社会で発生した事象を理解するには、高頻度で使われていた単語について分析する必要があると考え、登場する単語の頻度分析を行った。

まずは、全期間における単語の登場頻度の順位とその分布を調べた。また、期間ごとの単語の出現頻度を可視化するワードクラウドによる分析を行った。ワードクラウドではその時々における単語の出現頻度の断面はわかるものの、時間変化を可視化することはできない。そこで、いくつかの単語を抜き出して単語の発生数の時系列を可視化した。これにより、どのような事象がいつ発生して、どれくらいの期間継続的に注目されていたのか理解することができる。コロナや新型肺炎、感染症など関連語句を比較し、当時発生していた事象と照らし合わせることにより社会情勢の推移を数値やグラフで可視化することができた。

今回行った結果を以下にまとめる。

・極性値によって、定量的に社会全体のポジティブ、ネガティブの極性の代表値を求め、その推移をグラフによって可視化
・ワードクラウドによって、ある時期に高頻度で登場する語句の可視化
・単語使用数のグラフによって、語句ごとの使用状況の変化の可視化

これらにより、ある時刻を指定した時に起きていた事象とある事象の注目度の時間変化を捉えることができる。

最後に今後の展望を述べる。現在は、単純に単語の使用頻度からどのような事象が注目されていたかを推測した。これだけでも大まかな傾向はつかむことができる。しかし、より詳細な内容の分析のためには、同じ見出しの中で同時に使用される単語の関連性分析をする必要があるだろう。例えば、同じ単語であっても結びつく語によってその使われ方や意味が変わってくることも考えられる。そのため、ワードクラウドによる可視化の次のステップとして、重み付き無向グラフによる単語の関連性の分析などが考えられる。これは今後の課題である。

〈参考文献〉

熊本忠彦，田中克己．“Web ニュース記事からの喜怒哀楽抽出．”*情報処理学会研究報告自然言語処理 (NL)* 2005.1(2004-NL-165)（2005）：15－20.

厚生労働省，新型コロナウィルスについて オープンデータ，https://www.mhlw.go.jp/stf/covid-19/open-data.html,（2021年12月３日確認）

国立感染症研究所，現場からの概況：ダイアモンドプリンセス号における COVID−19症例，https://www.niid.go.jp/niid/ja/diseases/ka/corona-virus/2019-ncov/2484-idsc/9410-covid-dp-01.html,（2021年12月３日確認）

小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一. 意見抽出のための評価表現の収集．自然言語処理（2005），Vol. 12，No. 3，pp. 203－222.

佐藤 翔輔，林 春男，井上 和治，西野 隆博，（2009）「ウェブニュースに見る災害・危機における社会的側面の時系列展開の可視化」，可視化情報学会論文集，7 号，29巻.

佐藤敏紀，橋本泰一，and 奥村学.“単語分かち書き用辞書生成システム NEologd の運用―文書分類を例にして.”*研究報告自然言語処理 (NL)* 2016.15（2016）：1 －14.

佐藤敏紀，橋本泰一，and 奥村学.“単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討.”*言語処理学会第23回年次大会発表論文集*（2017）：875－878.

高津弘明，安藤涼太，松山洋一，& 小林哲則．（2021）．ニュース記事のタイトルと文の系列に対する感情分析．In 人工知能学会全国大会論文集 第 35 回全国大会（2021）（pp. 2Yin508-2Yin508）．一般社団法人 人工知能学会.

高野憲悟，& 萩原将文．（2011）．感情関連語を用いた感情推定法の提案とニュースサイトのアクセス解析への応用．In *日本知能情報ファジィ学会 ファジィシステム シンポジウム 講演論文集 第27回ファジィシステムシンポジウム*（pp. 150－150）．日本知能情報ファジィ学会.

東山昌彦，乾健太郎，松本裕治，述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第14回年次大会論文集（2008），pp. 584－587.

宮原淳．（2019）．地域ニュースの需要―オンラインの人気ランキング分析から．岐阜聖徳学園大学紀要．外国語学部編，58，95－103.

Kudo Taku, Kaoru Yamamoto, and Yuji Matsumoto. “Applying conditional random fields to Japanese morphological analysis.” *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.

Sato, Toshinori. “Neologism dictionary based on the language resources on the Web for Mecab.” *GitHub*（2015）.

Zipf, George Kingsley, The Psycho-Biology of Language, Boston-Cambridge Mass.: Houghton Mifflin（1935）.

Deciphering Social Change from Time Series Data in Access Rankings of Web News Articles

MICHIKOSHI Shugo
MARUNO Yuki

〈Abstract〉

With the spread of the Internet, it is becoming more common to read newspaper articles on websites in addition to traditional print newspapers. One of the features of newspapers on the Web is that they may include access rankings of news articles. Since the access ranking allows us to observe the articles that attract readers' attention, it is expected to analyze the data for understanding the social situation by extracting socially noteworthy events.

The purpose of this study is to examine how to easily understand the social situation by visualizing the information obtained from the access ranking of web articles. As a result of the verification using the headlines of web newspaper articles from 2017 to 2021, we found that the visualization is interpretable and easy to understand the meaning by analyzing the sentiment using polar dictionary and displaying the frequency of word usage using wordcloud.

Key words：Text mining, Sentiment analysis, Wordcloud