感情極性値を用いた Web ニュースの特徴語に対する評判分析
論文要旨

論文題目 | 感情極性値を用いたWebニュースの特徴語に対する評判分析
---|---

本研究では、Webニュースの本文を一ヶ月ごとの文書に分割し、特徴語とそれに関連する単語を用いて評判分析を行う。Web上で配信されているニュースは膨大な量であり、どのようなニュースが配信されているかを把握することは難しい。また、多くの文の集合である文書を対象にする場合は情報量が増え、さらに複雑化するので文書の性質を抽出することはより困難になる。

そこで本研究では、Web上で配信されているニュースの本文をテキストデータとして扱い、ニュースデータから一ヶ月ごとの評判を求める。具体的には、ニュースの本文を一ヶ月ごとに一つの文書として分割し、各月の特徴語を求め、それに関連する単語を用いて評判分析を行う。特徴語の決定にはTF-IDF値を用いる。これにより、他の月と比べたときに、特徴的な単語が抽出できる。また、評判分析の際には、感情極性値と呼ばれる単語の客観的な評判を数値化したものを利用する。

分析の結果から、特徴語がその月の特徴を表している単語だということがわかった。また、各月の評判は特徴語とその共起語の評判に大きく影響することがわかった。一方で、複数の月から同じ特徴語が抽出された場合でも、他の特徴語の影響を受け、それぞれの月の評判が異なることがわかった。
目次

第1章 はじめに................................................................. 1
  1.1 研究背景............................................................... 1
  1.2 研究動機............................................................... 2

第2章 先行研究............................................................. 3
  2.1 先行研究の概要 ..................................................... 3
  2.2 先行研究の課題と本研究の方向性 ......................... 3

第3章 関連知識............................................................ 5
  3.1 ニュースデータ ..................................................... 5
  3.2 形態素解析 ......................................................... 5
  3.3 感情極性値 ......................................................... 5
  3.4 単語の品詞 ......................................................... 6

第4章 分析手順.......................................................... 8
  4.1 単語の頻度の算出 ............................................... 8
  4.2 特徴語の抽出 ..................................................... 8
  4.3 共起語分析 ....................................................... 9
  4.4 感情極性値との対応 ............................................. 9

第5章 分析結果.......................................................... 10
  5.1 頻度分析結果 .................................................... 10
  5.2 特徴語抽出結果 ................................................. 11
  5.3 共起語分析結果 ............................................... 14
  5.4 感情分析結果 .................................................... 16

第6章 おわりに .......................................................... 17

参考文献 ......................................................................... 18

謝辞 .............................................................................. 19
第1章 はじめに

1.1 研究背景

近年では、インターネットや情報機器の発達に伴って、あらゆるものが電子化されている。特にテキストデータは急激な増加傾向にある。図1は企業が電子的に受信するデータ量の推移を表したグラフである[1]。このデータの対象となる業種は製造、建設、電力、ガス、水道、商業、金融、不動産、運輸、情報通信、サービスの9部門である。またデータの種類として、図2のように分類されている[1]。テキスト、音声、画像、動画といったデータの形式を横軸に取り、データの特性を縦軸に取った表である。こういったデータ流通量の増加により、企業や個人が多くの情報を発信するようになり、それに伴いそのデータを対象とした研究も盛んで行なわれている。そういった研究に企業も注目しており、自社の製品やサービスに対する評価を分析し、製品やサービスを改善する取り組みも行われている。たとえば、コールセンターに寄せられる意見を分類し、顧客がどのような意見や感想を持っているのかを分析することで、製品マニュアルを改善することにもつながる。

図1 企業のデータ流通量の推移
図2 ビッグデータ流通量の計量対象データ[1]

1.2 研究動機
インターネット上のデータの流通量が増えているということは1.1で述べた。これらのデータは発信者の考えや思いが反映されると考えられる。特にテキストデータでは、発信者の書き方によって読み手側が感じる印象が異なる。これらの評判は主観的なものであり、一意的に決めることは難しいとされてきた。また1つの単語や文に対してだけでなく多くの文の集合である文書を分析対象とするとき、その文書内で特徴的な単語がその文書全体の評判を左右すると思われる。この評判を知ることが出来れば、書き手が文章に込めた感情を読み手が正確に受け取ることができ、また書き手側も語弊を避けて文章を書くようになると考えられる。

そこで本研究では、ニューステキストデータから各月を一つの文書とし、月ごとに特徴的な語を求め、その月の評判を共起語の感情極性値を用いて求める。感情極性値は先行研究において機械的に作成された単語の評判を数値化したものであり、客観的に単語の評判を知ることができる。また特徴的な語であるということとは、その月の特徴を表すと考えられる。共起語はノードに対して似た性質を持つ単語なので、共起語を用いることで、特徴語の評判を求めることができる。特徴語の評判を求めることにより、各月の評判を決めることができると考えられる。
第2章 先行研究

本章では、本研究に関連する研究を紹介する。これらの研究は、Web上のテキストデータから実際の物事に対する評判を求めている。評判を抽出する研究はさまざまなものがあるが、本研究と関連の深い研究を以下に挙げる。

2.1 先行研究の概要

評判分析の研究は、J.Bollenらの研究を含む。この研究ではTwitter上のテキストデータと株価変動の相関関係を求め、株価推定を行っている。Opinion FinderとGPOMSを用いて、positive/negativeに振り分けを行う。Opinion Finderは2,718のpositiveな単語と4,912のnegativeな単語の計7,630語から形成された辞書を持つソフトである。GPOMSはGoogleウェブサイトから抽出した約1兆の単語を利用し、N-gram解析から作られた964語を収録した辞書である。Opinion FinderとGPOMSによって抽出された時系列データを平均値0、分散1の形に規格化し分析をしている。その結果86.7%という予測精度を出した。

WEB上のテキストデータから実社会における、評判を抽出することは十分に可能だと思われる。また緒方らの研究ではTF-IDF値と共感因子を用いて特徴語を抽出し、WEB上のテキストから人物評価を行っている。現実に存在する人に対しての評判を抽出しており、使用するデータによって実験結果が大きく異なるという点を述べている。使用するデータや指標は異なるが、客観的な評判の抽出に関する研究が進められれていることがわかる。これらの研究では身近で手に入れられるデータから、実際の事象についての評判を得ることができるということが共通点として挙げられる。

2.2 先行研究の課題と本研究の方向性

近年、テキストデータからの評判抽出に関する研究は盛んになっている。2.1で挙げた二つの研究は評判を抽出することに成功している。しかし、J.Bollenらの研究では7,630語から構築された辞書を使用しており、対象となる語彙数が少ないという問題がある。緒方らの研究では、評判を数値化した指標を扱っておらず、客観的に把握することが難しいという問題がある。またこれらの研究では単語同士の関係性を考慮に入れていない。評判を知るうえで対象となる単語と関わりの深い単語の評判を知ることも必要である。関わりの深い単語とその評判を知ることで、その単語の評判を正確に知ることができる。

そこで本研究では、単語同士の関係性を考慮に入れた評判分析を行う。ここで
は三年間のニュースデータを一ヶ月ごとに区切り、各月の評判を抽出する。各月の特徴語を選び、特徴的な語と関わりの深い単語を用いて評判分析を行う。特徴語はTF-IDF値を用いて抽出をする。また、関わりの深い単語にはコロケーションを用いて共起語を求める。
第 3 章 関連知識

本章では、本研究で用いる分析に関する知識について述べる。本研究では、ニュース記事の本文をデータとして扱っており、文章を単語に分割する手順と分割した後の単語の扱いと感情極性値について節ごとに述べる。

3.1 ニュースデータ

本研究で扱うニュースデータは、インターネット上で配信されている日本語ニュースの本文テキストを収集したものである。今回利用した配信サイトは表 1 に示している。またデータの形式は配信サイト名、配信サイトへのリンク、ニュースの配信日、ニュースのタイトル、ニュースの本文である。今回使用するのはニュースの配信日とニュースの本文のみである。ニュースデータの期間は 2012 年 1 月 1 日から 2014 年 12 月 31 日の 1,389,861 件である。

3.2 形態素解析

形態素とはこれ以上分けることのできない最小単位という意味である。文章を形態素解析するということは意味が存在する最小の単位に区切ることであり、品詞ごとに文章を区切り、単語を抽出する。また今回形態素解析に用いるソフトは MeCab である。MeCab は ChaSen を基に開発されたオープンソースの形態素解析ソフトであり、解析精度は ChaSen と同程度で、解析速度は平均して 3~4 倍の速さである。MeCab の特徴として、語彙や辞書、コーパスに依存しない汎用的な設計であるということが挙げられる。MeCab を用いて形態素解析を行うと、品詞、品詞細分類、活用形、活用型、原形、読み、発音の情報が出力される。本研究では、単語とその品詞、品詞細分類の情報のみを扱う。

3.3 感情極性値

感情極性値とは、単語の評判を表す数値である。ある単語に対して一つずつ割り振られている数値である。この値は -1 〜 1 の間の数値を取る。値が大きいほどポジティブな印象を持ち、小さいほどネガティブな印象を持つ。つまり、感情極性値が 0 より大きな単語はポジティブな印象を持ち、0 未満の単語はネガティブな印象を持つ。また、感情極性値の絶対値の大きさによってポジティブ/ネガティブ度合いが強くなることを示している。感情極性値は高村の研究[4]で機械的に算出されている。この研究では単語の感情極性を電子のスピンとみなして、語彙文、シソ
ラス及びコーパスで構築された語彙ネットワークをスピン系モデルとして用いている。さらに平均場近似を用いて近似確率分布関数を計算し、単語の感情極性値を算出している。この研究により、55135 単語に感情極性値が割り振られており、単語感情極性表として公開されている。この方法で感情極性値が割り振られた単語に対して、WordNet に収録されている語彙との比較をする実験を行ったところ、約 3000 語の単語を種とした場合には約 90%の正解率を示している。単語の評価を表す指標は複数あるが、感情極性値は高い精度と言える正解率を出ししており、この感情極性値を用いた評価分析に関する研究も行われている。

3.4 単語の品詞

収集したニューステキストを形態素解析した時に、各単語に品詞が割り振られる。この品詞は MeCab 内の辞書[5]に基づくものであり、53 種類に分けられる。その中で、今回使うのは表 2 の 10 個の品詞の種類を持つ単語である。使用品詞の選定の基準は、感情極性値を持つ単語の品詞に限定する。感情極性値を持たない単語も多く存在するので、品詞を限定することで、テキストのクリーニングを行うことができる。本研究におけるテキストのクリーニングとは、どのような文章でも出現する語を取り除くことである。対象となる語の例として句読点や記号、数詞などが挙げられる。こういった単語は文章の内容に関わらず出現するので、頻度が高く文書や他の単語に対する影響力が低い。これらの単語は数も多いので、分析を行う際の処理時間や結果の出力に影響を及ぼす。テキストのクリーニングを行うことにより、そういった分析の手順を大幅に削減することができる。
### 表1 ニュース配信サイト一覧

<table>
<thead>
<tr>
<th>配信元</th>
</tr>
</thead>
<tbody>
<tr>
<td>ロイター</td>
</tr>
<tr>
<td>毎日新聞</td>
</tr>
<tr>
<td>産経ニュース</td>
</tr>
<tr>
<td>朝日新聞デジタル</td>
</tr>
<tr>
<td>YOMIURI ONLINE</td>
</tr>
<tr>
<td>Yahoo!ニュース</td>
</tr>
<tr>
<td>NHK ニュース</td>
</tr>
<tr>
<td>J-CAST ニュース</td>
</tr>
<tr>
<td>goo ニュース</td>
</tr>
<tr>
<td>CNET JAPAN</td>
</tr>
</tbody>
</table>

### 表2 使用単語の品詞一覧

<table>
<thead>
<tr>
<th>品詞</th>
<th>品詞細分類</th>
</tr>
</thead>
<tbody>
<tr>
<td>名詞</td>
<td>固有名詞</td>
</tr>
<tr>
<td></td>
<td>一般</td>
</tr>
<tr>
<td>動詞</td>
<td>自立</td>
</tr>
<tr>
<td>形容詞</td>
<td>自立</td>
</tr>
<tr>
<td>副詞</td>
<td>一般</td>
</tr>
</tbody>
</table>
第4章 分析手順

本章では、本研究に必要な分析の手順を述べる。3.1で述べたニュースデータに対しても、3.2から3.4で述べた関連知識を用いて分析を行う。

4.1 単語の頻度の算出

ニュースデータを一ヶ月ごとの文書に区切り、MeCabを用いてすべての文書に形態素解析を行う。その後、各月の単語の出現頻度を求める。このとき、品詞を3.3で述べたものに絞り、頻度が10未満の単語は影響力が弱いとし、取り除いた。また各文書と単語の頻度行列を作成する。今回は使用する3年間のニュースデータなので、それを一ヶ月ごとに分割した36文書を使用する。

4.2 特徴語の抽出

特徴語の抽出の指標としてTF-IDF値を使用する。TF-IDF値とは、文書内の単語の頻度を表すTerm Frequencyの略であるTFと、単語が出現する文書数の逆数の逆数を対数で取ったInverse Document Frequencyの略であるIDFをかけあわせたものである。TF、IDFはそれぞれ以下の式(1)、(2)で表される。IDFの対数は重み付けの役割を果たしているので、対数の底は任意に決めることができる。今回は対数の底は2とする。ある語が対象となる全ての文書内に現れていればIDFは0となるので特徴的でないとわかる。今回、4.1で求めた全ての単語に対して、TF-IDF値を求め、その値の上位3単語を特徴として選ぶ。式(1)における$n_{i,j}$は単語$i$の文書$j$における頻度であり、分母は文書$j$におけるすべての単語の頻度の合計である。

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$IDF_i = \log_2 \frac{|D|}{df_i} \quad (2)$$

$|D|$は総文書数、$df_i$は単語$i$を含む文書数である。このため多くの文書に出現する語のIDF値は下がる。
4.3 共起語分析

共起とはノードと呼ばれる特定の語に対して、深い関わりがあるということを意味している。自然言語処理の分野では文章内で近くに出現する確率が高いほどより共起しているといえる[6]。また共起をしているかどうかの指標として $T$ 値と $MI$ 値がある。これらの値は以下の式(1),(2)で表される。コーパス言語学では $T>1.65, MI>1.58$ を満たしていると共起しているといえる[7]。

$$T = \frac{\text{共起頻度} - \text{共起期待値}}{\sqrt{\text{共起頻度}}} \tag{3}$$

$$MI = \log_2 \frac{\text{共起頻度}}{\text{共起期待値}} \tag{4}$$

$T$ 値は共起頻度と共起期待値の差を取り、共起頻度の平方根で調整した値である。$MI$の特徴として、低頻度であるが共起関係がある語を抽出できるという特徴がある。この 2 つの条件を満たしている語を共起語とし、4.2 で求めた特徴語の共起語を求める。

4.4 感情極性値との対応

高村らの研究[4]において感情極性値を求める際に、関連の深い単語は近い感情極性値を持つという仮定で算出している。4.3 で述べたが、共起語は近い性質を持つ単語と共起しやすいのので、共起語はノードに対して近い感情極性値を持つと考えられる。例えば犯罪のニュースでは「殺人」と「逮捕」などがよく共起しており、近い感情極性値を持つ。ネガティブな単語に対してはネガティブな単語が共起しやすい。また逆に、ポジティブな単語に対してはポジティブな単語が共起しやすい傾向がある。特徴語は固有名詞や新語であることが多いので、感情極性値が先行研究によって与えられていないことがある。したがって 4.3 で抽出した特徴語の共起語を利用し、共起語の感情極性値を対応することで、似た性質を持つ特徴語の評価を知ることができる。また特徴語は文書を特徴づけるので、その文書の評価も知ることができる。今回の分析では、各月の評価を特徴語の共起語の頻度に感情極性値をかけあわせた値の平均値で評価する。
第5章 分析結果

5.1 頻度分析結果

3年分のニュースデータを形態素解析し、品詞によって絞りこんだものを1年ごとの頻度で降順に表したもののが以下の表3である。この表からわかるとおり毎年同様の単語が頻度上位となっている。これは月ごとの頻度分析をした場合も同様である。多種多様なニュースが配信されているが、日々変化していく出来事をニュースとして伝えているものが多い。名詞では「日本」や「東京」といった国名や地名が上位になるやすい。日本語のニュースであるため、「日本」という単語が最上位にランクインしていると思われる。また、他国名では「中国」が3年間ランクインしている。外交についてのニュースでの中国についてのものの割合が多いということがわかる。動詞では「行う」や「いう」が毎回上位に入っており、よく使われる単語は毎年変わらないことがわかる。形容詞や副詞は表現の種類の多さから上位に上がることは少ない。分析対象期間やニュースを配信する媒体によってこの傾向は異なると思われる。こういったニュースの特徴を掴むことが出来れば、どのような単語が頻度上位になるのかが、ある程度決まってくることがわかる。

表3 1年ごとの頻度上位単語

<table>
<thead>
<tr>
<th>順位</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>単語</td>
<td>頻度</td>
<td>単語</td>
</tr>
<tr>
<td>1</td>
<td>日本</td>
<td>172,289</td>
<td>日本</td>
</tr>
<tr>
<td>2</td>
<td>行う</td>
<td>118,927</td>
<td>東京</td>
</tr>
<tr>
<td>3</td>
<td>東京</td>
<td>99,494</td>
<td>行う</td>
</tr>
<tr>
<td>4</td>
<td>いう</td>
<td>91,694</td>
<td>いう</td>
</tr>
<tr>
<td>5</td>
<td>中国</td>
<td>88,957</td>
<td>ない</td>
</tr>
<tr>
<td>6</td>
<td>ない</td>
<td>87,286</td>
<td>中国</td>
</tr>
<tr>
<td>7</td>
<td>政府</td>
<td>83,776</td>
<td>政府</td>
</tr>
<tr>
<td>8</td>
<td>述べる</td>
<td>81,314</td>
<td>示す</td>
</tr>
<tr>
<td>9</td>
<td>述す</td>
<td>78,150</td>
<td>受ける</td>
</tr>
<tr>
<td>10</td>
<td>受ける</td>
<td>77,536</td>
<td>述べる</td>
</tr>
</tbody>
</table>
5.2 特徴語抽出結果

文書とする各月を横軸に取り、単語を縦軸に取った頻度行列を作成し、全ての文書の語に対して TF-IDF 値を求めた。3年間の各月の最上位の単語を表したもののが表4、表5、表6である。これらの語は特徴語と呼ばれる。また、各月の最も特徴的な単語とその TF-IDF 値を表したもののが表7である。これらの表を見るときのように、5.1で上位に入っていた単語はランクインしていない。多くの文書に出現する単語は IDF が小さくなるので、それに伴って TF-IDF 値もまた小さくなる。これにより、一般的な単語が除かれて文書を比較した際に特徴的と思われる単語が抽出できる。また、IDF の値が大きくても TF の値が小さい単語は特徴語として抽出されない。これらの表の単語は各月の特徴的な語を表しているので、その月独特の単語が上位として抽出される。抽出された特徴語は他の文書と比較したときに特徴的とされる単語であり、その月にどのようなニュースが配信されたかというのを知る指標になる。

表4において、「日食」と「金環」が5月の特徴語として1位と2位にランクインしている。金環日食の話題から抽出された特徴語だと思われる。このような一つの出来事から、特徴的な単語が複数抽出されることがある。金環日食のニュースが他の月に比べて特に特徴的であることがわかる。他の月の場合、別の話題から上位3位にランクインしているものもしく、同程度に特徴的な出来事が複数あったということがわかる。

<table>
<thead>
<tr>
<th>月</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>アイオワ</td>
<td>オウム真理教</td>
<td>ギングリッチ</td>
</tr>
<tr>
<td>2</td>
<td>冠動脈</td>
<td>真部</td>
<td>ヒューストン</td>
</tr>
<tr>
<td>3</td>
<td>インパクト</td>
<td>地滑り</td>
<td>弾力</td>
</tr>
<tr>
<td>4</td>
<td>森内</td>
<td>藤崎</td>
<td>亀岡</td>
</tr>
<tr>
<td>5</td>
<td>日食</td>
<td>金環</td>
<td>ホルムアルデヒド</td>
</tr>
<tr>
<td>6</td>
<td>菊地</td>
<td>克也</td>
<td>サリン</td>
</tr>
<tr>
<td>7</td>
<td>阿蘇</td>
<td>日田</td>
<td>小沢</td>
</tr>
<tr>
<td>8</td>
<td>反日</td>
<td>釣魚</td>
<td>台風</td>
</tr>
<tr>
<td>9</td>
<td>森口</td>
<td>角田</td>
<td>ドラム缶</td>
</tr>
<tr>
<td>10</td>
<td>ガザ</td>
<td>崖</td>
<td>ハマス</td>
</tr>
<tr>
<td>11</td>
<td>笹子</td>
<td>勘三郎</td>
<td>天井板</td>
</tr>
<tr>
<td>12</td>
<td>アルジェリア</td>
<td>日揮</td>
<td>人質</td>
</tr>
</tbody>
</table>
表 5 2013 年の特徴語

<table>
<thead>
<tr>
<th>月</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>隕石</td>
<td>グループホーム</td>
<td>気球</td>
</tr>
<tr>
<td>2</td>
<td>キプロス</td>
<td>ブエルトリコ</td>
<td>ドミニカ共和国</td>
</tr>
<tr>
<td>3</td>
<td>ボストン</td>
<td>淡路島</td>
<td>サッチャー</td>
</tr>
<tr>
<td>4</td>
<td>飯島</td>
<td>オクラホマ</td>
<td>勘</td>
</tr>
<tr>
<td>5</td>
<td>コンフェデレーションズ</td>
<td>都議</td>
<td>イスタンブール</td>
</tr>
<tr>
<td>6</td>
<td>モル</td>
<td>萩</td>
<td>クーデター</td>
</tr>
<tr>
<td>7</td>
<td>延岡学園</td>
<td>同胞</td>
<td>お盆</td>
</tr>
<tr>
<td>8</td>
<td>越谷</td>
<td>ブエノスアイレス</td>
<td>兵器</td>
</tr>
<tr>
<td>9</td>
<td>台風</td>
<td>元町</td>
<td>伊豆</td>
</tr>
<tr>
<td>10</td>
<td>レイテ島</td>
<td>防空</td>
<td>タクロバン</td>
</tr>
<tr>
<td>11</td>
<td>防空</td>
<td>マンデラ</td>
<td>猪瀬</td>
</tr>
<tr>
<td>12</td>
<td>細川</td>
<td>農薬</td>
<td>都知事</td>
</tr>
</tbody>
</table>

普段はあまり使われないが、季節によって使われるようになる単語もまた特徴語として抽出されるということがわかる。

多くの月で共通してランクインする単語は無いが、2014年では「ウクライナ」が3ヶ月に渡ってランクインしている。2014年2月に発生したウクライナ内戦に

表 6 2014年の特徴語

<table>
<thead>
<tr>
<th>月</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>大雪</td>
<td>ソチ</td>
<td>河内</td>
</tr>
<tr>
<td>2</td>
<td>ウクライナ</td>
<td>クリミア</td>
<td>クリミア半島</td>
</tr>
<tr>
<td>3</td>
<td>ウクライナ</td>
<td>小保</td>
<td>旅客船</td>
</tr>
<tr>
<td>4</td>
<td>ウクライナ</td>
<td>ウルムチ</td>
<td>炭鉱</td>
</tr>
<tr>
<td>5</td>
<td>コートジボワール</td>
<td>コロンビア</td>
<td>コスタリカ</td>
</tr>
<tr>
<td>6</td>
<td>ガザ</td>
<td>ベネッセ</td>
<td>ハマス</td>
</tr>
<tr>
<td>7</td>
<td>安佐南</td>
<td>安佐北</td>
<td>ガザ</td>
</tr>
<tr>
<td>8</td>
<td>代々木公園</td>
<td>蚊</td>
<td>伝染病</td>
</tr>
<tr>
<td>9</td>
<td>御嶽山</td>
<td>小渕</td>
<td>エボラ出血熱</td>
</tr>
<tr>
<td>10</td>
<td>高倉</td>
<td>サンゴ</td>
<td>エボラ出血熱</td>
</tr>
<tr>
<td>11</td>
<td>阿蘇</td>
<td>日田</td>
<td>小沢</td>
</tr>
<tr>
<td>12</td>
<td>海江田</td>
<td>東海林</td>
<td>シャーロット</td>
</tr>
</tbody>
</table>

表5から、2013年7月の特徴語には第3位に「お盆」がランクインしている。
ついてのニュースが、他の月に比べて多く配信されるようになっただことが原因と思われる。2月では「クリミア」もまたウクライナ内戦に関連する単語である。しかし、3月と4月では他にウクライナ内戦に関連する単語はランクインしていない。この結果から、ウクライナ内戦についてのニュースは徐々に減少傾向にあることがわかる。ウクライナ内戦のような長期的な出来事に対しては、長期的にニュースを配信する場合があり、IDF が低くなるが、特徴語として上位に挙がってくる。

表7から、最もTF-IDF値が高い月は2015年9月の「御嶽山」であり、最も低いのは2014年11月の「阿蘇」である。この二つのTF-IDF値の差は10,000以上あり、各月で最も特徴的であってもその特徴度合いが異なることがわかる。「御嶽山」が特に高い TF-IDF 値を示したのは2014年9月に発生した御嶽山の噴火が原因だと考えられる。御嶽山の噴火は、連日ニュースで取り上げられていた。今回使用したWebニュースデータにおいても、御嶽山の噴火についてのニュース記事が急激に増加したため、高い TF-IDF 値を取ったと思われる。

表7 各月のTF-IDF値と最上位単語

<table>
<thead>
<tr>
<th>月</th>
<th>単語</th>
<th>TF-IDF</th>
<th>2013 単語</th>
<th>TF-IDF</th>
<th>2014 単語</th>
<th>TF-IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>アイオワ</td>
<td>1,518.3</td>
<td>隕石</td>
<td>3,037.5</td>
<td>大雪</td>
<td>7,000.4</td>
</tr>
<tr>
<td>2</td>
<td>冠動脈</td>
<td>2,068.8</td>
<td>キプロス</td>
<td>8,952.7</td>
<td>ウクライナ</td>
<td>10,693.5</td>
</tr>
<tr>
<td>3</td>
<td>インパクト</td>
<td>1,723.4</td>
<td>ボストン</td>
<td>4,629.1</td>
<td>ウクライナ</td>
<td>5,043.1</td>
</tr>
<tr>
<td>4</td>
<td>森内</td>
<td>1,583.6</td>
<td>飯島</td>
<td>4,708.7</td>
<td>ウクライナ</td>
<td>3,697.3</td>
</tr>
<tr>
<td>5</td>
<td>日食</td>
<td>7,205.6</td>
<td>コンフェデレーションズ</td>
<td>2,662.1</td>
<td>コートジボワール</td>
<td>6,339.9</td>
</tr>
<tr>
<td>6</td>
<td>菊地</td>
<td>7,159.5</td>
<td>モル</td>
<td>3,291.3</td>
<td>ガザ</td>
<td>6,864.6</td>
</tr>
<tr>
<td>7</td>
<td>阿蘇</td>
<td>6,091.3</td>
<td>延岡学園</td>
<td>2,969.9</td>
<td>安佐南</td>
<td>12,358.9</td>
</tr>
<tr>
<td>8</td>
<td>反日</td>
<td>5,937.6</td>
<td>越谷</td>
<td>4,339.4</td>
<td>代々木公園</td>
<td>8,208.6</td>
</tr>
<tr>
<td>9</td>
<td>森口</td>
<td>2,576.7</td>
<td>台風</td>
<td>5,490.1</td>
<td>御嶽山</td>
<td>12,958.0</td>
</tr>
<tr>
<td>10</td>
<td>ガザ</td>
<td>5,847.4</td>
<td>レイテ島</td>
<td>4,626.6</td>
<td>高倉</td>
<td>4,949.8</td>
</tr>
<tr>
<td>11</td>
<td>笹子</td>
<td>3,348.3</td>
<td>防空</td>
<td>4,333.0</td>
<td>阿蘇</td>
<td>1,314.0</td>
</tr>
<tr>
<td>12</td>
<td>アルジェリア</td>
<td>5,252.7</td>
<td>細川</td>
<td>5,626.5</td>
<td>海江田</td>
<td>1,444.1</td>
</tr>
</tbody>
</table>

13
5.3 共起語分析結果

各月のTF-IDF値上位100単語の共起語を全て求めた。その中でも4.2で述べた式(3)と式(4)の$T$値と$MI$値の条件を満たすもののみを抽出した。これにより、共起しているといえる単語のみが抽出され、各月の特徴語と関連がある単語を得た。また共起関係にある単語においても、頻度が高いほど関連があるとし、影響力が強いと仮定する。2012年から2014年の特徴語に対して、共起頻度が上位3単語の共起語とその感情極性値を以下の表8,表9,表10に示す。

### 表8 2012年の共起語

<table>
<thead>
<tr>
<th>月</th>
<th>1</th>
<th>極性値</th>
<th>2</th>
<th>極性値</th>
<th>3</th>
<th>極性値</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>急落</td>
<td>−0.14826</td>
<td>苦戦</td>
<td>−0.98795</td>
<td>健闘</td>
<td>0.98184</td>
</tr>
<tr>
<td>2</td>
<td>栄養</td>
<td>0.90013</td>
<td>狭心症</td>
<td>−0.41663</td>
<td>病院</td>
<td>−0.61506</td>
</tr>
<tr>
<td>3</td>
<td>OK</td>
<td>0.98960</td>
<td>インパクト</td>
<td>0.94008</td>
<td>働け</td>
<td>−0.13829</td>
</tr>
<tr>
<td>4</td>
<td>後手</td>
<td>−0.43742</td>
<td>今</td>
<td>0.00531</td>
<td>今日</td>
<td>0.26450</td>
</tr>
<tr>
<td>5</td>
<td>ブロポーズ</td>
<td>0.09598</td>
<td>過去</td>
<td>−0.60585</td>
<td>楽しめる</td>
<td>0.97700</td>
</tr>
<tr>
<td>6</td>
<td>意気込む</td>
<td>0.99216</td>
<td>改称</td>
<td>−0.04842</td>
<td>疑い</td>
<td>−0.98869</td>
</tr>
<tr>
<td>7</td>
<td>一例</td>
<td>0.97285</td>
<td>勧告</td>
<td>0.08348</td>
<td>頃</td>
<td>0.12579</td>
</tr>
<tr>
<td>8</td>
<td>スト</td>
<td>−0.27708</td>
<td>汚職</td>
<td>−0.98765</td>
<td>過剰</td>
<td>−0.55086</td>
</tr>
<tr>
<td>9</td>
<td>長</td>
<td>0.95732</td>
<td>岳</td>
<td>−0.04596</td>
<td>娘</td>
<td>−0.04654</td>
</tr>
<tr>
<td>10</td>
<td>支援</td>
<td>0.98973</td>
<td>支配</td>
<td>−0.72474</td>
<td>実効</td>
<td>0.07672</td>
</tr>
<tr>
<td>11</td>
<td>頑張る</td>
<td>0.98170</td>
<td>五</td>
<td>0.82389</td>
<td>順調</td>
<td>0.03269</td>
</tr>
<tr>
<td>12</td>
<td>事件</td>
<td>−0.85996</td>
<td>処罰</td>
<td>−0.99480</td>
<td>体罰</td>
<td>−0.97889</td>
</tr>
</tbody>
</table>

### 表9 2013年の共起語

<table>
<thead>
<tr>
<th>月</th>
<th>1</th>
<th>極性値</th>
<th>2</th>
<th>極性値</th>
<th>3</th>
<th>極性値</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ミス</td>
<td>−0.73235</td>
<td>よく</td>
<td>0.92187</td>
<td>育てる</td>
<td>0.93694</td>
</tr>
<tr>
<td>2</td>
<td>殺人</td>
<td>−0.99128</td>
<td>容疑</td>
<td>−0.48305</td>
<td>改革</td>
<td>−0.02998</td>
</tr>
<tr>
<td>3</td>
<td>インフルエンザ</td>
<td>−0.56074</td>
<td>ウイルス</td>
<td>−0.83709</td>
<td>家</td>
<td>0.00246</td>
</tr>
<tr>
<td>4</td>
<td>新</td>
<td>0.97812</td>
<td>元年</td>
<td>−0.04349</td>
<td>違反</td>
<td>−0.62758</td>
</tr>
<tr>
<td>5</td>
<td>コメント</td>
<td>0.08346</td>
<td>ライバル</td>
<td>0.98852</td>
<td>閉鎖</td>
<td>−0.61323</td>
</tr>
<tr>
<td>6</td>
<td>抗争</td>
<td>−0.58342</td>
<td>当時</td>
<td>0.01511</td>
<td>無血</td>
<td>0.04975</td>
</tr>
<tr>
<td>7</td>
<td>ショック</td>
<td>−0.98592</td>
<td>混雑</td>
<td>−0.99195</td>
<td>祝日</td>
<td>0.98085</td>
</tr>
<tr>
<td>8</td>
<td>テロリスト</td>
<td>−0.25928</td>
<td>ミサイル</td>
<td>−0.30798</td>
<td>家</td>
<td>0.00246</td>
</tr>
<tr>
<td>9</td>
<td>報</td>
<td>0.12295</td>
<td>死亡</td>
<td>−0.76567</td>
<td>事故</td>
<td>−0.99705</td>
</tr>
<tr>
<td>10</td>
<td>反論</td>
<td>−0.39223</td>
<td>非難</td>
<td>−0.99823</td>
<td>問題</td>
<td>−0.71676</td>
</tr>
<tr>
<td>11</td>
<td>トップ</td>
<td>0.91168</td>
<td>茶人</td>
<td>−0.03356</td>
<td>武将</td>
<td>0.31267</td>
</tr>
<tr>
<td>12</td>
<td>強風</td>
<td>−0.53281</td>
<td>恐れ</td>
<td>−0.99904</td>
<td>警戒</td>
<td>−0.69215</td>
</tr>
</tbody>
</table>
表 10 2014 年の共起語

<table>
<thead>
<tr>
<th>月</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>単語</td>
<td>極性値</td>
<td>単語</td>
<td>極性値</td>
</tr>
<tr>
<td>1</td>
<td>悪化</td>
<td>-0.99176</td>
<td>意欲</td>
</tr>
<tr>
<td>2</td>
<td>サポート</td>
<td>0.97885</td>
<td>リスク</td>
</tr>
<tr>
<td>3</td>
<td>インフレ</td>
<td>-0.03573</td>
<td>エネルギー</td>
</tr>
<tr>
<td>4</td>
<td>一番</td>
<td>0.96356</td>
<td>無料</td>
</tr>
<tr>
<td>5</td>
<td>テロ</td>
<td>-0.98767</td>
<td>緊張</td>
</tr>
<tr>
<td>6</td>
<td>エネルギー</td>
<td>0.92092</td>
<td>テロ</td>
</tr>
<tr>
<td>7</td>
<td>対立</td>
<td>-0.99388</td>
<td>過去</td>
</tr>
<tr>
<td>8</td>
<td>困難</td>
<td>-0.99556</td>
<td>混乱</td>
</tr>
<tr>
<td>9</td>
<td>リスク</td>
<td>-0.99065</td>
<td>援助</td>
</tr>
<tr>
<td>10</td>
<td>御嶽山</td>
<td>0.00000</td>
<td>県境</td>
</tr>
<tr>
<td>11</td>
<td>日田</td>
<td>0.00000</td>
<td>故郷</td>
</tr>
<tr>
<td>12</td>
<td>政権</td>
<td>0.00000</td>
<td>防衛</td>
</tr>
</tbody>
</table>

2013年3月の共起語を見てみると、「インフルエンザ」、「ウイルス」が上位になっている。この結果からインフルエンザに関するニュースが多く配信されたと思われる。上位100個の特徴語を対象にした共起語の集合であるが、共起頻度上位になっているものが同じ内容のニュースになっているとわかる。このことから2013年3月においてはインフルエンザウイルスに関するニュースがより特徴的なものだとわかる。

表10における極性値が0となっている単語については感情極性値を持たないということを表す。9月の特徴語である「御嶽山」が10月の共起語の頻度で1位にランクインしており、特徴語が変化しても御嶽山の噴火に関するニュースの配信が10月も続いていることがわかる。
5.4 感情分析結果

各月ごとに全ての共起語の感情極性値と共起頻度を掛け合わせ平均を取った。以下の図がその値の推移を表したものであり、横軸は一ヶ月ごとの時間帯を表していて、縦軸は感情極性値を表している。

図3からわかるように感情極性値がマイナスになっている月のほうが多い。プラスになっている月では共起語の極性値が大きいものが多く、逆にマイナスになっている月では共起語の極性値が小さくなる傾向にある。これは、似た性質の単語と共起しやすいという共起語の性質によるものである。極性値の近い単語同士が共起しやすいため、月ごとの極性値に大きな差が出たと思われる。

図3 各月の特徴語に対する共起語の平均感情極性値の推移
第6章 おわりに

本研究では、特徴語と共起語、感情極性値を組み合わせることによって文書の評判分析を行った。具体的には、各文書から特徴語を抽出し、その特徴語に対する共起語を求め、共起語の共起頻度と感情極性値を掛け合わせたものの平均をその文書の評判とした。共起語は似た性質の語と共起しやすいという特徴を利用したものである。また日々配信されるニュースデータの中で、個性とも言える特徴語を用いることで、文書ごとの評判の違いを数値的に表した。

これらの分析によって出た結果は、すでに5章で述べた。今回は文書が一ヶ月区切りで期間が連続だったので、各月の評判の推移という形で表した。5章で述べたが、影響力の強い語に引かれてその月の評判が定まっているように感じた。単純な頻度ではなく、TF-IDF値が高い特徴語を用いることで低頻度の語が大きな影響を与えるとわかった。

客観的な評判を分析によって抽出することを目的として本研究を進めてきた。今回は共起語の性質を利用して評判を求めたが、この指標を一つのみでは説得力に欠ける部分もあると考えられる。また、今回求めた月ごとの評判に対して、他の研究との比較や正答率の検証を行っていないので、どの程度正確なものかも不明である。今後は多角的に見て評価ができるように新たな指標の提案を行いたい。また評判分析を行っている先行研究とも比較をし、正答率等の比較をする必要がある。
参考文献

[1] 総務省：情報通信白書 企業のデータ流通量の推計結果 平成26年版


[3] 緒方進, 池田真司, 牟田高信:Web 上のテキスト情報を用いた

[4] 高村大也, 乾孝司, 奥村学, スピンモデルによる単語の感情極性抽出, 情報

[5] MeCab:品詞IDの定義,
   (2016年1月22日確認)

[6] 石川慎一郎,言語コーパスからのコロケーション検出の手法,統計数理研究所