平成 27 年度　卒業論文
2016 年 2 月 2 日

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

法政大学　理工学部　経営システム工学科

経営数理工学研究室

12X4117　中本泰斗

指導教員　五島洋行教授

学科名	経営システム工	学籍番号	12X4117
申請者氏名		中本　泰斗	
指導教員氏名		五島　洋行	

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ヵ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　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に行われている.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判を分析し,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.

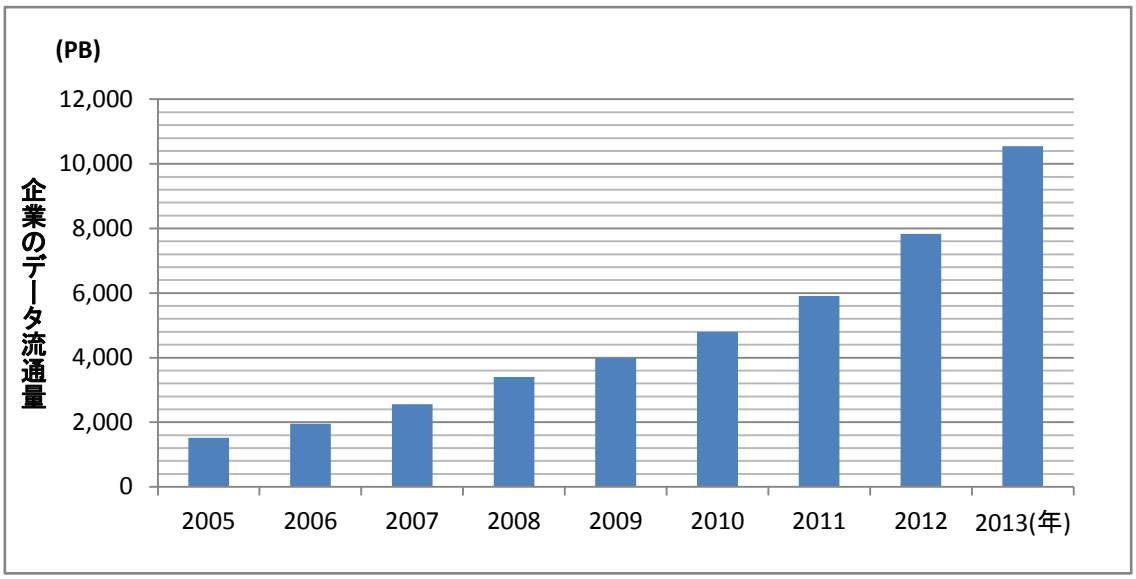

図１　企業のデータ流通量の推移

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可能になり,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[2]などがある. 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 が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[3]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ことを述べている.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てき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[2]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[3]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,言語や辞書,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 の間の数値を取り,値が大きいほどポジティブな印象をもち,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 より大きい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,語釈文,シソ

ーラス,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.単語の評判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ており,この感情極性値を用いた評判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

配信元
ロイター
毎日新聞
産経ニュース
朝日新聞デジタル
YOMIURI ONLINE
Yahoo!ニュース
NHK ニュース
J-CAST ニュース
goo ニュース
CNET JAPAN

表 2　使用単語の品詞一覧

品詞	品詞細分類
名詞	固有名詞
	一般
	代名詞
	非自立
	副詞可能
動詞	自立
	非自立
形容詞	自立
	非自立
副詞	一般

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,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.今回使用するのは 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割合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.と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.今回は,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 単語を特徴語とする.(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.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$IDF_i = \log_2 \frac{|D|}{df_i} \tag{2}$$

|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する.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判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,その文書の評判も知ることができる.今回の分析では,各月の評判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りやすい.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

順位	2012		2013		2014	
	単語	頻度	単語	頻度	単語	頻度
1	日本	172,289	日本	211,572	日本	182,231
2	行う	118,927	東京	136,419	行う	120,274
3	東京	99,494	行う	134,132	東京	117,787
4	いう	91,694	いう	121,172	いう	99,869
5	中国	88,957	ない	112,660	ない	96,160
6	ない	87,286	中国	107,877	選手	86,207
7	政府	83,776	政府	104,685	中国	85,841
8	述べる	81,314	示す	91,797	政府	83,850
9	示す	78,150	受ける	91,374	示す	78,491
10	受ける	77,536	述べる	89,084	述べる	75,823

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 値を表したものが表 7 である.これらの表を見るとわかるように,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.このように一つの出来事から,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多く,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.

表 4　2012 年の特徴語

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

表 5　2013 年の特徴語

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

表 6　2014 年の特徴語

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

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.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

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

月	2012		2013		2014	
	単語	TF-IDF	単語	TF-IDF	単語	TF-IDF
1	アイオワ	1,518.3	隕石	3,037.5	大雪	7,000.4
2	冠動脈	2,068.8	キプロス	8,952.7	ウクライナ	10,693.5
3	インパクト	1,723.4	ボストン	4,629.1	ウクライナ	5,043.1
4	森内	1,583.6	飯島	4,708.7	ウクライナ	3,697.3
5	日食	7,205.6	コンフェデレーションズ	2,662.1	コートジボワール	6,339.9
6	菊地	7,159.5	モル	3,291.3	ガザ	6,864.6
7	阿蘇	6,091.3	延岡学園	2,969.9	安佐南	12,358.9
8	反日	5,937.6	越谷	4,339.4	代々木公園	8,208.6
9	森口	2,576.7	台風	5,490.1	御嶽山	12,958.0
10	ガザ	5,847.4	レイテ島	4,626.6	高倉	4,949.8
11	笹子	3,348.3	防空	4,333.0	阿蘇	1,314.0
12	アルジェリア	5,252.7	細川	5,626.5	海江田	1,444.1

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

月	1		2		3	
	単語	極性値	単語	極性値	単語	極性値
1	急落	−0.14826	苦戦	−0.98795	健闘	0.98184
2	栄養	0.90013	狭心症	−0.41663	病院	−0.61506
3	OK	0.98960	インパクト	0.94008	値上げ	−0.13829
4	後手	−0.43742	今	0.00531	今日	0.26450
5	プロポーズ	0.09598	過去	−0.60585	楽しめる	0.97780
6	意気込む	0.99216	改称	−0.04842	疑い	−0.98869
7	一角	0.97285	勧告	0.08348	頃	0.12295
8	スト	−0.27708	汚職	−0.98765	過剰	−0.55086
9	長	0.95732	岳	−0.04596	娘	−0.04654
10	支援	0.98973	支配	−0.72474	実効	0.07672
11	頑張る	0.98170	五	0.82389	順調	0.03269
12	事件	−0.85996	処罰	−0.99480	体罰	−0.97889

表 9　2013 年の共起語

月	1		2		3	
	単語	極性値	単語	極性値	単語	極性値
1	ミス	−0.73235	よく	0.92187	育てる	0.93694
2	殺人	−0.99128	容疑	−0.48305	改革	−0.02998
3	インフルエンザ	−0.56074	ウイルス	−0.83709	家	0.00246
4	勲	0.97812	元年	−0.04349	違反	−0.62758
5	コメント	0.08346	ライバル	0.98852	閉鎖	−0.61323
6	抗争	−0.58942	当時	0.01511	無血	0.04975
7	ショック	−0.98592	混雑	−0.99195	祝日	0.98085
8	テロリスト	−0.25928	ミサイル	−0.30798	家	0.00246
9	頃	0.12295	死亡	−0.76567	事故	−0.99705
10	反論	−0.39223	非難	−0.99823	問題	−0.71676
11	トップ	0.91168	茶人	−0.03356	武将	0.31267
12	強風	−0.53281	恐れ	−0.99904	警戒	−0.69215

表 10　2014 年の共起語

月	1		2		3	
	単語	極性値	単語	極性値	単語	極性値
1	悪化	−0.99176	意欲	0.96545	下落	−0.99119
2	サポート	0.97885	リスク	−0.99065	悪化	−0.99176
3	インフレ	−0.03573	エネルギー	0.92092	タイミング	0.08500
4	一番	0.96356	無料	0.11757	ゲリラ	−0.22008
5	テロ	−0.98767	緊張	−0.56186	支配	−0.72474
6	エネルギー	0.92092	テロ	−0.98767	トップ	0.91168
7	対立	−0.99388	過去	−0.60585	不安	−0.82962
8	困難	−0.99556	混乱	−0.72734	支援	0.98973
9	リスク	−0.99065	援助	0.99544	強化	0.97029
10	御嶽山	0.00000	県境	0.00000	噴火	−0.54568
11	日田	0.00000	故郷	0.00000	祝賀	0.99250
12	政権	0.00000	防衛	0.04668	問題	−0.71676

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.

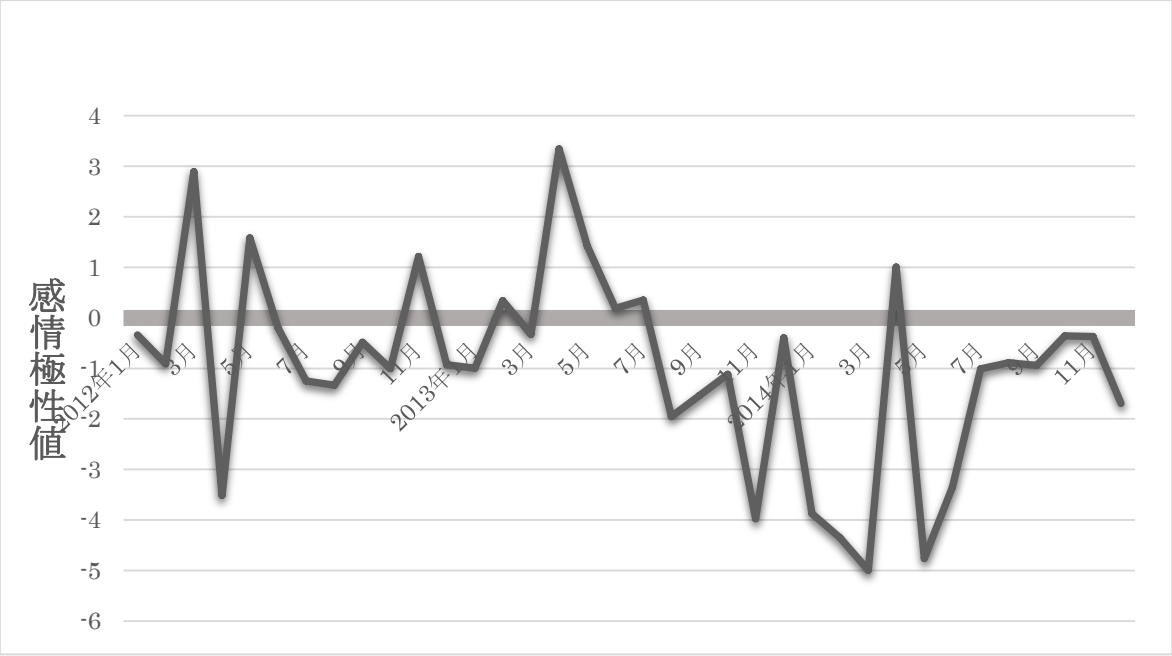

図 3　各月の特徴語に対する共起語の平均感情極性値の推移

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 章ですでに述べたが,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
http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h26/html/nc131220.html （2016 年 1 月 22 日確認）

[2] J.Bollen, H.Mao, X-J.Zeng:Twitter mood predicts the stock market, Journal of Computation Science, Vol.2, No.1, pp.1–8 2010

[3] 緒方進, 池田真司, 牟田高信, 木本勝敏:Web 上のテキスト情報を用いた人物評価手法, 情報処理学会研究報告自然言語処理 Vol.2005, No.1, pp.9–14, 2005.

[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, 情報処理学会論文誌ジャーナル, Vol.47, No.2 pp.627–637, 2006.

[5] MeCab:品詞 ID の定義,
http://mecab.googlecode.com/svn/trunk/mecab/doc/posid.html
(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,統計数理研究所共同研究レポート,Vol.1,No.190,pp.1–14,2006.

[7] 石田基広:R によるテキストマイニング入門, 森北出版 (2008)