センチメント分析に基づくニュース記事の信憑性判断支援

松本　好史† 　張　　建偉† 　河合由起子† 　中島　伸介† 　熊本　忠彦††

田中　克己†††

† 京都産業大学　〒 603-8555 京都市北区上賀茂本山
†† 千葉工業大学　〒 275-0016 千葉県習志野市津田沼 2-17-1
††† 京都大学　〒 606-8501 京都市左京区吉田本町
E-mail: †{g738444,zjw,kawai,nakajima}@cc.kyoto-su.ac.jp, ††kumamoto@net.it-chiba.ac.jp,
†††tanaka@dl.kuis.kyoto-u.ac.jp

あらまし　Web の普及によりニュース記事を発信することが容易になったが，発信された記事の中には発信サイト自身の立場に基づいて書かれた記事も存在し，そのような記事によって偏った情報を得てしまう危険性がある．そこで本稿では，センチメント分析に基づくニュース記事の信憑性判断を支援するシステムを提案する．本システムでは，記事のセンチメントと発信サイトの過去のセンチメントとの差異，および発信サイトのセンチメントと他のサイトのセンチメントとの差異の 2 点をグラフ化してユーザに提示することができる．本システムを利用することで，ユーザに記事やサイトの立場を客観的に把握させることができ，ニュース記事の信憑性判断を支援することができる．

キーワード　センチメント，ニュース記事，信憑性

1. はじめに

近年，web の普及により web 上で情報を収集・発信することが容易になった．そのため，情報収集の手段の一つとして web が活用されるようになった．しかし，容易に情報が発信できてしまうために偏った情報も多く発信されてしまっており，情報の質の低下が問題となっている．

特にニュース記事の場合，情報の質の低下は深刻な問題である．なぜならニュース記事は，起こった出来事を素早く正確に伝えるものであり，人々はニュース記事から得た情報を元に物事を考える，あるいは行動するからである．したがって，発信サイトにとって利益になるニュース記事のみを発信する，あるいは発信サイトの利益になるようにニュース記事を編集するといった事態が発生すると，ニュース記事を読んだ人の判断を誤らせ損失を与えてしまう危険性がある．

そのような危険を防ぐためには，ある事柄に対して異なる立場のニュース記事を比較しながら情報を得ることが望ましい．しかし，ニュース記事や発信サイトがどのような立場をとっているかを客観的に判断し，それらと異なる立場のニュース記事やサイトをユーザに提示するようなシステムはない．

本稿では，センチメントを用いてニュース記事や発信サイトの立場をユーザに視覚的に提示するシステムを提案する．なお，センチメントとはニュース記事から抽出した単語を元に感情軸ごとに算出した値のことであり，現在使用している感情軸は「楽しい⇔悲しい」,「嬉しい⇔怒り」,「のどか⇔緊迫」の 3 軸である [1]．本システムでは，まずユーザが信憑性を判断したいニュース記事を選択する．次に，システムはキーワードと関連語をユーザに提示するので，ユーザは興味のあるキーワードと関連語を選択する．なお，キーワードとはニュース記事内から抽出した重要語であり，関連語は関連記事から抽出した重要語である．その後，システムは選択したニュース記事のセンチメントと発信サイトの過去のセンチメントとの差異，および発信サイトのセンチメントと他のサイトのセンチメントとの差異の 2 点をユーザに提示する．

例として，読売新聞の「「機密共有」裏目の米省庁…ウィキリークス流出」というタイトルのニュース記事 (図 1) においてキーワードを「ウィキリークス」，関連語を「情報」としたときの「楽しい⇔悲しい」という感情軸におけるこのニュース記事のセンチメントと読売新聞の「ウィキリークス」「情報」に関する過去のセンチメントとの差異を図 2 に示す．読売新聞の過去のセンチメントと比べて，この記事がより悲しいセンチメントを持つことがわかる．また,「ウィキリークス」「情報」に関して，読売新聞と過去のセンチメントが似ている新聞社として抽出された時事通信，読売新聞と過去のセンチメントが異なる新聞社として抽出された毎日新聞，および読売新聞のセンチメントを図 3 に示す．本システムを利用することでユーザに異なる感情を持つニュース記事を容易に発見させることができ，信憑性判断の支援を行うことができる．

2. 関連研究

2.1 センチメント分析に関する研究

Turney らはレビューを相互情報量を基に「推薦する」「推薦しない」の 2 つのカテゴリーに分類する手法を提案した [2]．Pang らは映画レビューの主観的な部分を抜き出しテキストカテゴリゼーション技術を適用することにより,「満足」「不満」に分類する手法を提案した [3]．Liu らはセンチメントを製品セールスパフォーマンスの予測に利用するための感情認識のモデルを紹介した [4]．しかし，これらの方法はセンチメントがポジティ

「機密共有」裏目の米省庁…ウィキリークス流出
民間の内部告発サイト「ウィキリークス」が、機密扱いを含む米政府の外交公電を連日公開しており、全世界に波紋が広がっている。
米外交を揺るがす前代未聞の情報流出劇はなぜ起きたのか。
公電２５万通
今回ウィキリークスが入手し、公開するとしているのは、ワシントンの国務省と世界２７４か所に展開する
米大使館・領事館が１９６６年１２月～２０１０年２月に交わした公電２５万１２８７通。
最も秘密性が高い「最高機密（ｔｏｐｓｅｃｒｅｔ）」文書は含まれていないが、
それに準じる「機密（ｓｅｃｒｅｔ）」扱いが1万５６５２通、
さらにその下に位置づけられる「秘密（ｃｏｎｆｉｄｅｎｔｉａｌ）」も１０万１７４８通含まれている。
このうち、在日米大使館と国務省の間で交わされた公電は５６９７通で、うち２２７通が機密扱い。
いずれもまだ、サイト上には出ていないが、米軍普天間飛行場の移設問題などを巡る微妙なやり取りが含まれている可能性もある。
ウィキリークスのサイト上では１１月２８日に約２３０通、２９日にも約４０通が公開されたほか、
事前にウィキリークスから公電を一部提供されたニューヨーク・タイムズ紙など欧米の報道機関も内容を報道している。
だれが盗んだ？
では、だれがこれらの公電を「盗み出した」（クリントン国務長官）のか。
疑惑の渦中にいるのが、ウィキリークスに軍事情報を流したとして７月に起訴された、ブラッドリー・マニング陸軍上等兵。
イラク駐留当時に政府・軍関係者が省庁間の秘密文書を共有する情報ネットワークにアクセスしては、
音楽ＣＤに見せかけた記録可能なＣＤに情報をダウンロードしていたとされる人物だ。
この情報システムは、「機密ＩＰルーターネットワーク（ＳＩＰＲＮｅｔ）」と呼ばれる。
米政府は、０１年の同時テロ後、政府機関同士の情報共有が十分でなかったとの反省から、
国務省を含む複数の省庁がネットワーク上に乗り入れ、機密情報を共有するシステムを構築していた。
ワシントン・ポスト紙によると、このシステムを利用出来る政府・軍関係者は５０万～６０万人。
マニング上等兵もその１人だった。
後手に回る防止策
今回の事態を受け、国務省は、このシステムから自省のコンピューターを一時切り離した。
国防総省も、機密情報をＣＤなど持ち出し可能な記録媒体に書き込めないようにすると同時に、
だれが機密文書を閲覧しているかを監視するなどの対策を講じている。
しかし、国防総省のホイットマン報道官は、閲覧監視システムは現在、同省が扱う機密文書全体の６割程度しか
カバー出来ていない状況だと説明している。（ワシントン黒瀬悦成）
（2010 年 12 月 1 日 読売新聞）

図１　分析対象である記事１

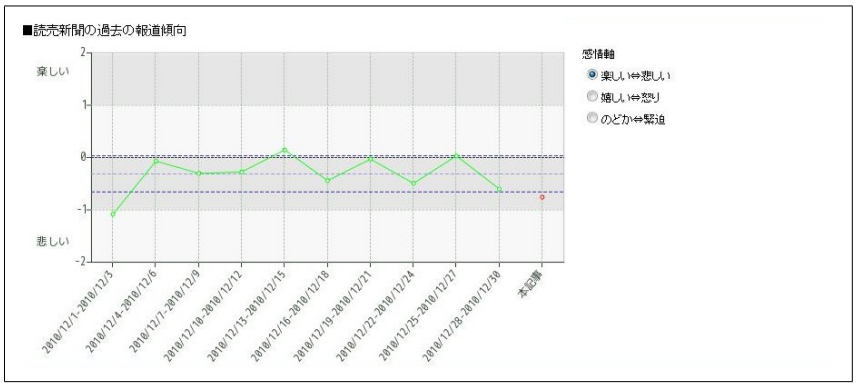

図２　記事１に対する記事のセンチメントと発信サイトの過去のセンチメントとの差異

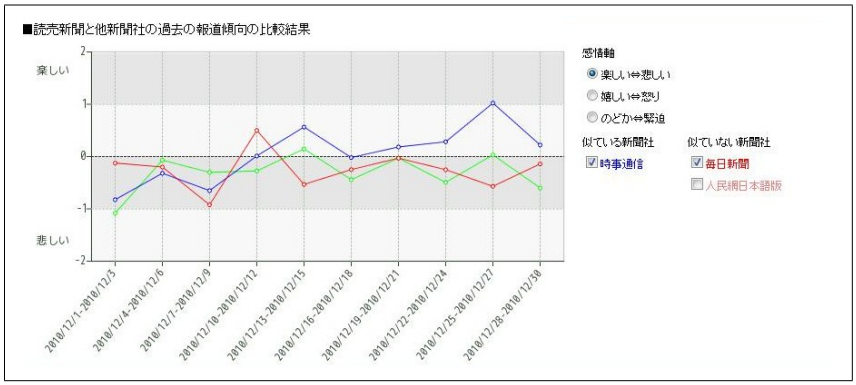

図３　記事１に対する発信サイトのセンチメントと他のサイトのセンチメントとの差異

ブであるかネガティブであるかを考慮するだけである．我々は地域ごとの違いにより，どのようなセンチメントに基づいて情報が発信されているかを可視化するセンチメントマップシステムを提案した [5]．ポジティブとネガティブなセンチメントだけでなく，より人間の感情に近いとされる感情モデルに基づき 4 次元のセンチメント（「明るい⇔暗い」,「承認⇔拒否」,「緩和⇔緊張」,「怒り⇔恐れ」）を分析した．

本稿では，ニュース記事には適当と思われる「楽しい⇔悲しい」,「嬉しい⇔怒り」,「のどか⇔緊迫」の 3 軸を用いてセンチメント分析を行い，センチメント分析の結果を用いてニュース記事の信憑性判断を支援するシステムを提案している．

2.2 情報の信憑性に関する研究

情報の信憑性に関しては多くの論文が発表されており，様々な手法で研究が行われている．特徴的な語を抽出し記事のテーマにおけるそれらの語の妥当性から文章の特異度を算出する手法 [6]，ページのリンクや被リンクの数あるいは質などから信憑性を算出する方法 [7] [8]，発信者の特徴を捉えそこから信憑性を算出する手法 [9] などがある．

本稿では，センチメントの異なる記事やサイトをユーザに提示しニュース記事の偏りを防ぐことで，ニュース記事の信憑性判断を支援している．

3. システムの概要

本システムの処理は，オフライン処理とオンライン処理に分けることができる (図 4)．

オフライン処理では，まず新聞社サイトからニュース記事を取得する．次に取得した記事に対して形態素解析を行い単語を抽出し，$tf \cdot idf$ 値を算出する．その後，抽出した単語と感情辞書を用いてニュース記事のセンチメントを算出する．

オンライン処理では，まずユーザがニュース記事を選択する．次に，選択した記事を元にキーワードと関連語をユーザに提示する．ユーザがキーワードと関連語を選択すると，選択したニュース記事のセンチメントと発信サイトの過去のセンチメントとの差異，および発信サイトのセンチメントと他のサイトのセンチメントとの差異の 2 点をユーザに提示する．

4. オフライン処理

オフライン処理では，ニュース記事の取得，$tf \cdot idf$ 値の算出，センチメントの算出の 3 つの処理を行う．

4.1 ニュース記事の取得

オフライン処理では，まずニュース記事の取得を行う．なお，現在は国内 15 社，国外 10 社の計 25 社の新聞社サイトからニュース記事の取得を行っている．次に取得したニュース記事から HTML タグなどを取り除きタイトルと本文を抽出し，データベースに登録する．

4.2 $tf \cdot idf$ 値の算出

$tf \cdot idf$ 値の算出では，まず取得したタイトルと本文に対して Juman [10] を用いて形態素解析を行い単語を抽出し，各単語について tf 値を算出する．なお，タイトルに使われている単語に関しては tf 値を 3 倍にする．次に，データベースから

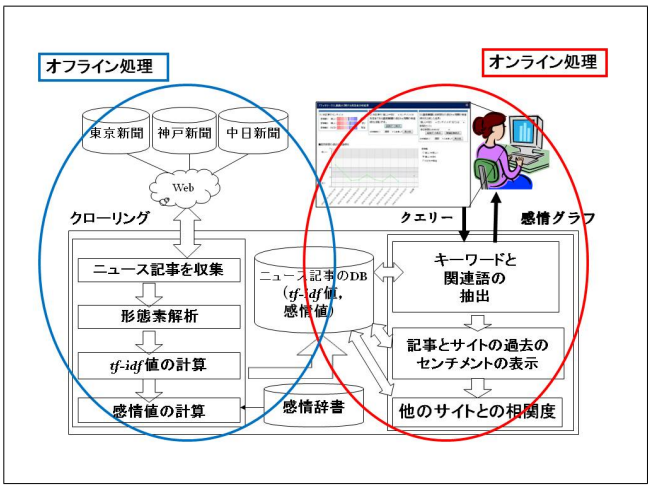

図 4　システムの概要

タイトルまたは記事に各単語を含む記事の総数を取得し，idf 値を算出する．最後に，tf 値と idf 値から $tf \cdot idf$ 値を算出し，データベースに登録する．なお，記事 Pi で抽出された単語 w の出現回数を $N(w, Pi)$，Pi で抽出された全単語の出現回数を $N(Pi)$，全ドキュメント数を N，w を含むドキュメント数を $N(w)$ とすると，Pi における w の $tf \cdot idf$ 値 $tf \cdot idf(w, Pi)$ は以下の式で求められる．

$$tf \cdot idf(w, P_i) = \frac{N(w, P_i)}{N(P_i)} \cdot log\frac{N}{N(w)}$$

4.3 センチメントの算出

センチメントの算出には感情辞書を使用する．感情辞書の作成方法を説明すると，まず感情軸を構成する感情語群を設定する (表 1)．次に感情語を含む記事を抽出し記事に含まれる感情語群 IW_L に属する感情語と感情語群 IW_R に属する感情語の数を比較し，IW_L の数が多い記事の集合を S_L(記事数を N_L)，IW_R の数が多い記事の集合を S_R(記事数を N_R) とする．このとき，ある単語 w の記事集合 S_L における出現頻度を N_L(w)，記事集合 S_R における出現頻度を $N_R(w)$ とすると，それぞれの補正済み条件付確率は，

$$P_L(w) = \frac{N_L(w)}{N_L}$$
$$P_R(w) = \frac{N_R(w)}{N_R}$$

と表される．この $P_L(w)$ と $P_R(w)$ を用いて，単語 w のセンチメント $s(w)$ を次のような式で表す．

$$s(w) = \frac{P_L * weight_L}{P_L(w) * weight_L + P_R(w) * weight_R}$$

$$weight_L = log_{10}N_L$$
$$weight_R = log_{10}N_R$$

センチメント $s(w)$ は 0〜1 の値をとる．1 に近い値は「楽し

い，嬉しい，のどか」という感情を表し，0 に近い値は「悲しい，怒り，緊迫」という感情を表す．感情辞書の一例を表 2 に示す．この例では,「初受賞」という単語の「楽しい⇔悲しい」という感情軸のセンチメントは 0.862 であり,「楽しい」という感情を表す.「偽装」という単語の「嬉しい⇔怒り」という感情軸のセンチメントは 0.075 であり,「怒り」という感情を表す.記事のセンチメントは，記事に出現した各単語のセンチメントの平均で算出される．

表 1　感情軸と感情語群

感情軸	感情語
楽しい (L)	楽しい，楽しむ，楽しみだ，楽しげだ (L)
⇔悲しい (R)	悲しい，悲しむ，悲しみだ，悲しげだ (R)
嬉しい (L)	嬉しい，喜ばしい，喜ぶ (L)
⇔怒り (R)	怒る，憤る，激怒する (R)
のどか (L)	のどかだ，和やかだ，素朴だ，安心だ (L)
⇔緊迫 (R)	緊迫する，不気味だ，不安だ，恐れる (R)

表 2　感情辞書一例

単語 1 ⇔ 0	感情軸 1 楽しい⇔悲しい	感情軸 2 嬉しい⇔怒り	感情軸 3 のどか⇔緊迫
初受賞	0.862	1.000	0.808
クッキング	1.000	0.653	0.881
ひなまつり	0.847	1.000	0.977
偽装	0.245	0.075	0.297
死刑だ	0.013	0.028	0.000
拘束する	0.059	0.103	0.000

5. オンライン処理

オンライン処理では，キーワードと関連語の抽出，サイトの過去のセンチメントの表示，他のサイトとの相関度の表示の 3 つの処理を行う．

5.1 キーワードと関連語の抽出

オンライン処理では，まずユーザが信憑性を判断したいニュース記事を選択する．次に，ユーザが選択した記事からタイトルと本文を取得し，$tf \cdot idf$ 値の高い 5 単語をキーワードとする．また，指定した期間内であり，かついずれかのキーワードを含む全記事から，$tf \cdot idf$ 値の和の高い語から順にキーワードと重複しないように 5 単語を抽出し関連語とする．その後，キーワードと関連語をそれぞれユーザに提示し，ユーザがそれらの中から分析したい対象を選択する．

5.2 記事とサイトの過去のセンチメントの表示

選択したキーワードと関連語におけるサイト内の過去のセンチメントを算出し，時間軸ごとにグラフ化して表示する．また，過去のセンチメントの平均値と標準偏差を求める．選択した記事のセンチメントが標準偏差の範囲内に入っていれば，記事とサイトのセンチメントが同じであるとし，標準偏差の範囲外であれば，記事とサイトのセンチメントが異なるとする．なお，分析期間を n 個の期間 t_i に分け，期間 t_i 内のユーザが選択したキーワードや関連語を含む記事のセンチメントの平均を $s(t_i)$ とすると，サイトのセンチメントの平均値 $\overline{s}$ と標準偏差 σ は以下の式で求められる．

$$\overline{s} = \frac{1}{n}\sum_{i=1}^{n} s(t_i)$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s(t_i) - \overline{s})^2}$$

5.3 他のサイトとの相関度の表示

まず，キーワードと関連語における各サイトのセンチメントを算出する．次に他のサイトとの相関係数を算出する．なお，分析期間を n 個の期間 t_i に分け，期間 t_i 内のサイト A の記事のセンチメントを $s_A(t_i)$，サイト B の記事のセンチメントを $s_B(t_i)$ としたときのサイト A とサイト B の相関係数 ρ は以下の式で求められる．

$$\rho(A,B) = \frac{\sum_{i=1}^{n}(s_A(t_i) - \overline{s_A}) * (s_B(t_i) - \overline{s_B})}{\sqrt{\sum_{i=1}^{n}(s_A(t_i) - \overline{s_A})^2} * \sqrt{\sum_{i=1}^{n}(s_B(t_i) - \overline{s_B})^2}}$$

そして，相関係数がしきい値 τ_1 以上のサイトをセンチメントが似ているサイト，しきい値 τ_2 以下のサイトを似ていないサイトとしユーザに提示する．

6. 実　　験

6.1 記事と発信サイトの過去のセンチメントの表示

読売新聞の「菅内閣支持率下落」というタイトルのニュース記事においてキーワードを「調査」，関連語を「内閣」としたときの記事とサイトの過去のセンチメントの表示を行った．なお，このニュース記事のセンチメントと読売新聞の過去のセンチメントとの「楽しい⇔悲しい」という感情軸における差異を図 5 に,「嬉しい⇔怒り」という感情軸における差異を図 6 に,「のどか⇔緊迫」という感情軸における差異を図 7 に示した．この結果，全ての感情軸においてこのニュース記事が持つセンチメントは従来の読売新聞のセンチメント傾向とは異なっていると提示された．したがって，このニュース記事を読む際にはこのニュース記事が本当に信頼できるかどうかを警戒しながら読む必要がある．

6.2 発信サイトと他のサイトのセンチメントの比較

先ほどと同条件で他のサイトとの相関度の表示を行った.「楽しい⇔悲しい」という感情軸において読売新聞と過去のセンチメントが似ているサイトは毎日新聞，異なっているサイトは CNN と提示された (図 8)．また,「嬉しい⇔怒り」という感情軸において似ているサイトは朝鮮日報日本語版，異なっているサイトは CNN(図 9),「のどか⇔緊迫」という感情軸において似ているサイトは朝鮮日報日本語版，異なっているサイトは人民網日本語版と提示された (図 10)．この結果からこの記事を読む際には CNN や人民網日本語版の記事と読み比べて信憑性を判断するのがよいといえる．

7. まとめと今後の課題

本システムでは，記事のセンチメントと発信サイトの過去の

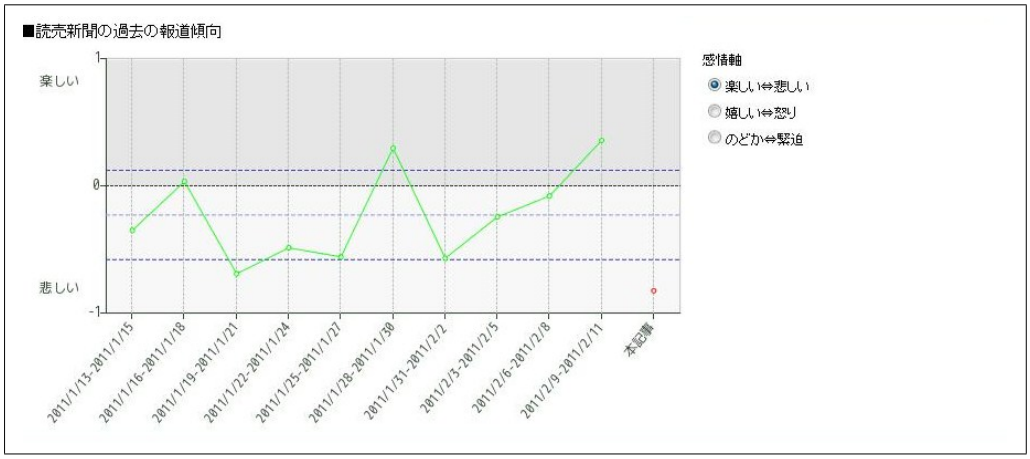

図 5　「楽しい⇔悲しい」という感情軸における記事のセンチメントと発信サイトの過去のセンチメントとの差異 (「調査」「内閣」に対する分析結果)

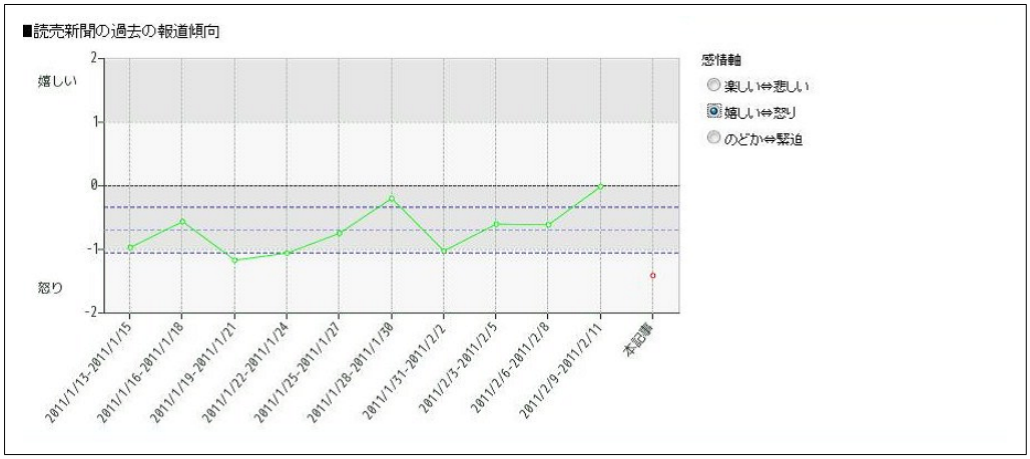

図 6　「嬉しい⇔怒り」という感情軸における記事のセンチメントと発信サイトの過去のセンチメントとの差異 (「調査」「内閣」に対する分析結果)

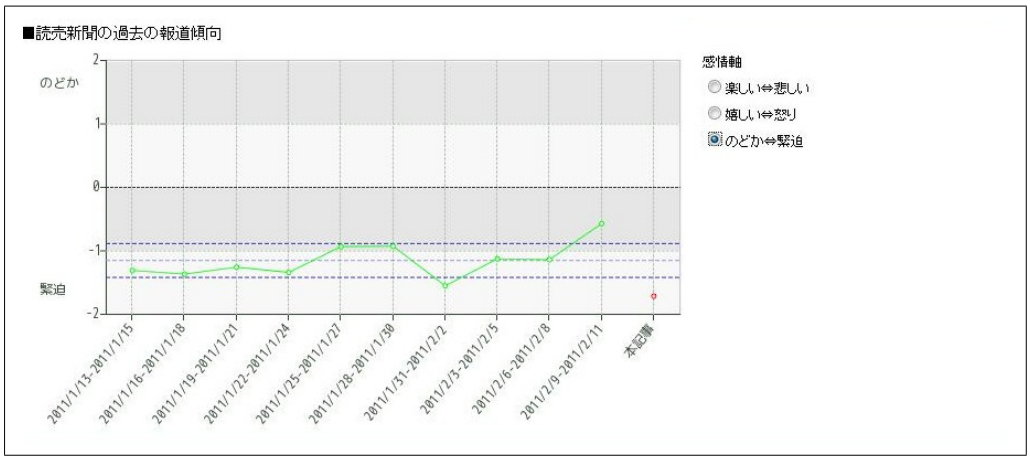

図 7　「のどか⇔緊迫」という感情軸における記事のセンチメントと発信サイトの過去のセンチメントとの差異 (「調査」「内閣」に対する分析結果)

センチメントとの差異，および発信サイトのセンチメントと他のサイトのセンチメントとの差異の 2 点をユーザに提示することができた．ユーザは記事やサイトの立場を客観的に把握することができ，比較対象となるサイトを容易に発見することが可能となりニュース記事の信憑性判断を支援することができた．

今後の課題としては，システムの精度の改良が挙げられる．また，センチメント分析に基づいた信頼度の算出方法も開発する予定である．

謝　　辞

この研究は，独立行政法人情報通信研究機構の高度通信・放送研究開発委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発 課題ア Web コンテンツ分析技術」の一環としてなされたものである．

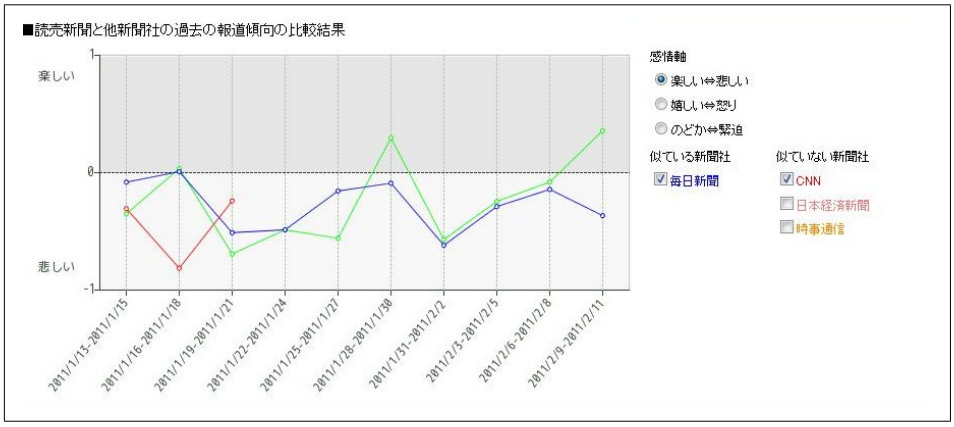

図 8　「楽しい⇔悲しい」という感情軸における発信サイトのセンチメントと他のサイトのセンチメントとの差異 (「調査」「内閣」に対する分析結果)

図 9　「嬉しい⇔怒り」という感情軸における発信サイトのセンチメントと他のサイトのセンチメントとの差異 (「調査」「内閣」に対する分析結果)

図 10　「のどか⇔緊迫」という感情軸における発信サイトのセンチメントと他のサイトのセンチメントとの差異 (「調査」「内閣」に対する分析結果)

文　　献

[1] 熊本忠彦，新聞記事を対象とする印象空間の構築，信学会第二種研資，Web インテリジェンスとインタラクション，WI2-2008-35, pp.47-52，2008.

[2] P. D. Turney,“ Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, ”In ACL 2002, pp. 417-424, 2002.

[3] B. Pang and L. Lee,“ A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, ”In ACL 2004, pp. 271-278, 2004.

[4] Y. Liu, X. Huang, A. An, and X. Yu,“ ARSA: A Sentiment-aware Model for Predicting Sales Performance Using Blogs, ” In SIGIR 2007, pp. 607-614, 2007.

[5] 張建偉，河合由起子，熊本忠彦，田中克己，地域性に基づく発信者の観点差異を可視化するセンチメントマップシステムの提案，情報処理学会論文誌：データベース，Vol. 3, No. 1 (TOD 45), pp. 38-48, 2010.

[6] 中林猛，湯本高行，新居学，高橋豊，Web を利用した語の詳細関係に基づく情報の信憑性判断支援，DEIM2010.

[7] 近藤浩之，手塚太郎，田中克己，地域的支持度に基づくウェブページの信頼性評価とオブジェクトレベル検索，DEWS2008.

[8] 井上雄介，太田学，脚注と参考文献を用いた Wikipedia 記事の信頼性評価の一手法，DEIM2010.

[9] 石田晋，馬強，吉川正俊，記述の主観性を考慮したニュース発信者の特徴分析とその応用，DEIM2010.

[10] Juman: http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html