情報処理学会論文誌　データベース　Vol. 3　No. 1　38–48 (Mar. 2010)

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

張　　建偉[†1]　河合　由起子[†1]
熊本　忠彦[†2]　田中　克己[†3]

近年，多くのニュースサイトが様々なサービスを提供するようになった．しかしながら，発信者の立場の違いによる観点の差異を発見・提示するための手法はあまり見られない．そこで，本研究では，地域ごとの違いにより，どのような観点（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システムを構築する．特に，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を分析する．また，地図を用いて，それらの観点の差異を地域ごとに提示する．本システムにより，ユーザはトピックに関する各ニュースサイトの記事を検索・閲覧できるだけでなく，地図の詳細度制御により，関東や関西，あるいは日本やアメリカといった地域性に基づいて，トピックに対する観点の相違を一元的に把握できる．本論文では，地域性に基づく発信者の観点相違を把握できるセンチメントマップシステムを提案するとともに，評価実験を行い，システムの有効性について検証する．

A Sentiment Map System for Visualizing the Distinction of Information Senders' Opinions Based on the Regionality

Jianwei Zhang,[†1] Yukiko Kawai,[†1]
Tadahiko Kumamoto[†2] and Katsumi Tanaka[†3]

Recently, an increasing number of news websites have come to provide various featured services. However, there has been little research on extracting and presenting the distinction of different news websites' viewpoints. Information senders (news websites) may report a same event with different opinions and sentiments. We develop a system for visualizing the distinction of opinions (sentiments) based on different regions. Specially, rather than positive and negative sentiments, we analyze more detailed sentiments with four dimensions, which are designed based on a sentiment model similar to human emotion. Moreover, we present the distinction of sentiments on a geographical map. Using our system, users can not only retrieve and browse news articles related to the concerned topic, but also grasp the sentiment distinction of different regions (e.g., Kanto versus Kansai, or Japan versus America) based on level-of-detail control of map scale. This paper describes a sentiment map system which can summarize the distinction of information senders' opinions based on the regionality, and verifies its effectiveness through several experiments.

1. はじめに

近年，Web の普及により多くの新聞社の記事を Web で自由に閲覧できるようになった[1)–3)]．これにともない，大量のニュース記事をユーザに効果的に提供するためのニュースポータルサイトが増えてきた[4),5)]．これらのサイトは，閲覧，キーワード検索，および個人化された様々なサービスを提供している．ユーザは，1 つのポータルサイトにアクセスすることによって欲しい情報を取得できるようになった．

著者らはこれまで複数のニュースサイトの記事をユーザの選好に基づいて推薦できるシステム MPV[6)] と MPV Plus[7)] を開発してきた．MPV では，ユーザの閲覧履歴からユーザの興味のあるキーワード（興味語）を抽出し，その興味語となるキーワードを基に記事を分類した後，興味語との関連度が高い記事を優先的に推薦できる．また，MPV Plus では，興味語との関連度だけでなく，各記事のセンチメント値とユーザが閲覧した記事からユーザのセンチメント値を抽出し，ユーザのセンチメント値と類似度の高い記事をランキングして推薦できる．これにより，ユーザは興味のある記事に容易にアクセスできるだけでなく，センチメントという新たな観点と類似する記事を閲覧することができる．たとえば，興味語が「イラク」の場合，各ニュースサイトのイラクの記事が自動分類され，さらに，このユーザが「イラクの明るいトピック」に興味があると判断した場合，ユーザは暗いトピックより明るいトピックの記事を優先的に閲覧できる．

しかし，興味やセンチメントといった個々のユーザの観点に合わせて個々の記事を推薦が

†1 京都産業大学
Kyoto Sangyo University
†2 千葉工業大学
Chiba Institute of Technology
†3 京都大学
Kyoto University

できる一方で，ドメインごとの観点をユーザは明確にとらえることができない．たとえば，任意の政党に関して情報発信される際，ドメインを新聞社とした場合，朝日新聞と読売新聞では観点が異なる場合がある．また，プロ野球の試合結果は，ドメインを関東と関西とした場合，関東圏の新聞と関西圏の新聞では異なった立場（勝者の立場/敗者の立場）から報じられる．さらには，ドメインを国ごととした場合，米国政府関連の記事は，親米か反米かによって国ごとで観点が異なる．

そこで，本研究では，このような各ドメインの観点の相違を，ユーザの地図の詳細度制御により抽出して時間軸にそって可視化するセンチメントマップシステムを提案する．本研究が着目する観点は，各記事の発信者（ニュースサイト，地方，国など）がそのトピックにいだくセンチメントとする．提案システムを用いることにより，ユーザは個々の記事を閲覧する前に，そのトピックに関する全体的なセンチメント傾向を新聞社レベル，地方レベルあるいは国レベルで把握できるようになり，そのようなセンチメント傾向を考慮したうえで，様々なセンチメントで書かれた記事を選択して閲覧できるようになる．特に，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」）を分析する．実際の新聞記事には，ポジティブ・ネガティブといったセンチメントだけでなく，様々なタイプのセンチメントが込められており，どのセンチメントに着目すべきかはトピックによっても異なる．たとえば，「プロ野球の試合結果」に関しては，「明るい⇔暗い」のセンチメントで記述される記事が多く，「承認⇔拒否」や「緩和⇔緊張」，「怒り⇔恐れ」といったセンチメントで記述される記事は少ない．一方，「事業仕分け」に関する記事では，「承認⇔拒否」あるいは「怒り⇔恐れ」といったセンチメントで記述される記事が多い．また，「景気回復」を「承認」するが，現状ではまだ「暗い」といった，ポジティブなセンチメントとネガティブなセンチメントを両方持つ記事も存在しうる．ユーザは，任意のトピックに関して，システムが提示する 4 次元のセンチメントを見ることにより，どのセンチメントに差異が現れているかを知ることができる．

図 1 は検索キーワード「中国」で生成されたセンチメントマップの例である．システムは，各地方ごとにニュース記事を検索し，記事のセンチメント値を 1 日ごとに集計し，時間軸にそってセンチメントグラフを作成する．図 1 の各地域ごとのグラフでは，4 つのセンチメント値をポジティブとネガティブの 2 つのセンチメントにまとめた例を提示している．この各グラフの上にマウスを置くことで，4 つの詳細なセンチメントグラフを閲覧できる．また，ユーザが地図をズームインすると，各ニュースサイトのセンチメントグラフが提示される（**図 2**）．ズームアウトすると，国ごとのセンチメントグラフが提示される（**図 3**）．

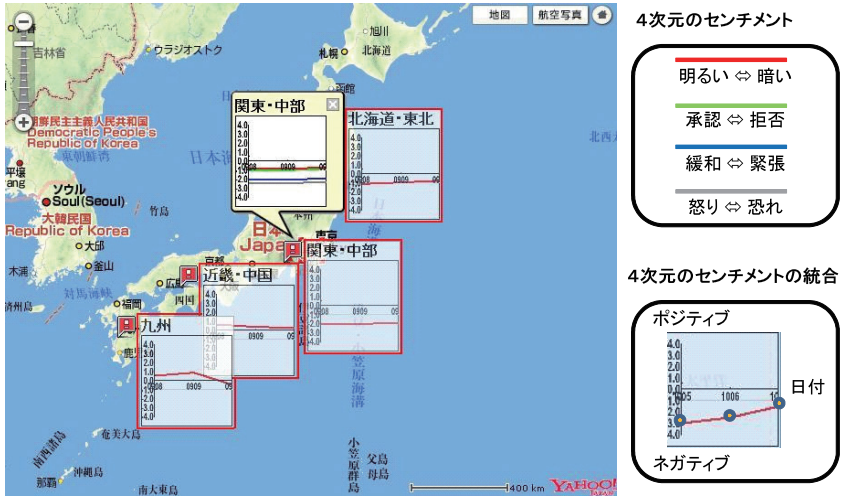

図 1　センチメントマップの例
Fig. 1　Example of a sentiment map.

以下，本論文では，地域性に基づくセンチメントの相違を抽出し可視化するシステムを提案するとともに，評価実験の結果に基づいて，システムの有効性を検討する．2 章では本システムの概要を述べる．3 章と 4 章ではシステムの事前処理と検索処理についてそれぞれの詳細を説明する．5 章では構築したシステムに関する実験結果について考察する．6 章では関連研究について述べ，7 章では最後にまとめとする．

2. システムの概要

提案システムの概要を**図 4** に示す．提案システムは，ニュース記事の収集および分析を行う事前処理と，記事の検索およびセンチメントマップの生成を行う検索処理からなる．

ユーザの検索前の事前処理として，システムは指定されたニュースサイトより記事を収集し，記事の形態素解析を行い単語を抽出し，その単語の $tf \cdot idf$ 値を算出する．また，これまでに著者らが開発してきたセンチメント辞書を用いて各記事に対する 4 次元のセンチメント値（「明るい⇔暗い」，「承認⇔拒否」，「緩和⇔緊張」，「怒り⇔恐れ」）を抽出しておく．

検索の際は，ユーザが任意のキーワードを入力すると，ニュースサイトごとにそのキー

図 2　ズームイン時のセンチメントマップ
Fig. 2　A sentiment map when zoomed in.

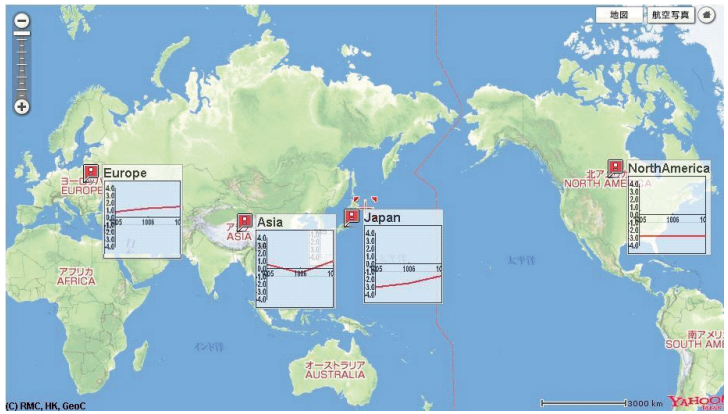

図 3　ズームアウト時のセンチメントマップ
Fig. 3　A sentiment map when zoomed out.

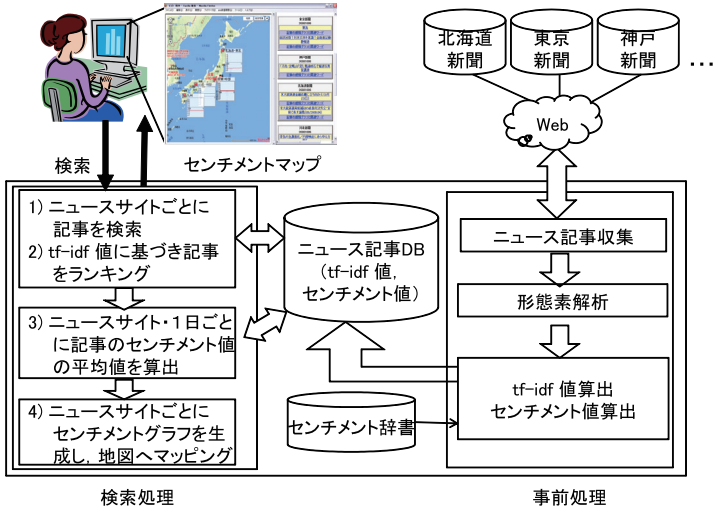

図 4　システムの概要
Fig. 4　System overview.

ワードを含む記事を検索し，各記事のキーワードに対する $tf \cdot idf$ 値を用いて記事を選ぶ．次に，サイトごとで 4 つのセンチメントに対して記事のセンチメント値の平均値を 1 日ごとに算出する．最後に，時間軸にそってセンチメントグラフを生成し，各ニュースサイトのグラフを地図上にマッピングし，ユーザに提示する．ユーザが地図をズームイン・ズームアウトした場合は，再度ドメインを決定し，新たなドメインのセンチメント平均値を再計算し，ユーザに提示する．

提案システムでは，ユーザは検索キーワードに関する各ニュースサイトの記事を検索・閲覧できるだけでなく，センチメントグラフを地図上にマッピングすることにより，センチメント差異が生じた地域を視覚的に把握できる．また，地図をズームイン・ズームアウトすることにより，地図の詳細度を容易に変えることができ，新聞社，地方，あるいは国のセンチメント傾向を一覧することができる．さらに，そのトピックに対するセンチメント値の時間的推移も把握できる．

3. 事 前 処 理

ユーザの検索以前に，センチメントマップシステムは以下のステップでニュース記事の収

集と分析を行う．

(1)　指定された複数のニュースサイトからニュース記事 P をクローリングする．

(2)　収集した P のテキストデータから HTML タグを除去する．

(3)　除去したテキストデータで形態素解析を行い，「動詞」，「名詞」，「副詞」，「形容詞」，「その他」の単語を抽出する．

(4)　P で抽出された単語 w の $tf \cdot idf$ 値を下記の式を用いて算出する．

$$tf \cdot idf(w, P) = \frac{C(w, P)}{C(P)} \times log \frac{N}{N(w)}$$

ただし，$C(w, P)$ は P で抽出された単語 w の出現回数，$C(P)$ は P で抽出された全単語の出現回数，N は複数のニュースサイトから収集した全記事数，$N(w)$ は単語 w を含む記事数である．

(5)　単語と単語のセンチメント値および重みの対応関係を示すセンチメント辞書を構築する．

(6)　センチメント辞書に基づき，各記事の 4 つのセンチメント値を算出する．

各ニュース記事のセンチメント値は，4 つのセンチメント尺度ごとに，記事に出現した単語のセンチメント値と重みをセンチメント辞書から取得し，計算式にあてはめることによって求められる．

以下，センチメント辞書を自動構築する手法（ステップ 5）ならびにニュース記事からセンチメント値を算出する手法（ステップ 6）について詳細を述べる．

3.1　センチメント辞書の構築

単語および記事のセンチメントは，心理学者 Plutchik の感情モデル[8] に基づき，「明るい⇔暗い」，「承認⇔拒否」，「緩和⇔緊張」，「怒り⇔恐れ」といった 4 つのセンチメント尺度とする．Plutchik の感情モデルでは，人間が持つすべての感情の基本となる 8 つの感情（基本感情あるいは一次感情）が提案されており，他の感情（二次感情）はこの基本感情を混合することで得られる，と定義されている．対極性（2 対 4 軸である点）と線形性（基本感情の組合せですべての感情を表現できる点）から，発信者のセンチメント差異の発見・提示という本研究の目的に対して適した計算モデルとなっている．

センチメント辞書を構築する際の基本的な考え方は，各センチメント尺度に対して，センチメント辞書において見出し語となる内容語と，シードとして与えたポジティブとネガティブな感情語群との共起関係を分析することである．4 つのセンチメント尺度とそれぞれの感情語群を**表 1** に示す．各センチメント尺度 $e \in \{a, b, c, d\}$ は，ポジティブな感情語群 e_1 とネガティブな感情語群 e_2 を持つ．たとえば，センチメント尺度「a：明るい⇔暗い」に対して，ポジティブな感情語群 $a_1 =$ { 明るい，うれしい，楽しい }，ネガティブな感情語群 $a_2 =$ { 暗い，悲しい，苦しい } となる．なお，感情語群は，対応する記事に現れやすく，そうでない記事に現れにくい感情・感覚形容詞を著者らの主観に基づいて決定した結果である．

センチメント辞書は日経新聞全文記事データベース（1990〜2001 年版の 200 万強の記事）を解析することにより構築された．**表 2** はセンチメント辞書の例であり，内容語 w と各センチメント尺度における内容語のセンチメント値 $S_e(w)$ と重み $M_e(w)$ の対応関係を表す．たとえば，内容語「蘇生」のセンチメント尺度「a：明るい⇔暗い」におけるセンチメント値は 0.91 であり，重みは 0.464 である．単語「蘇生」が「明るい」センチメントに偏るこ

表 1　センチメント尺度と感情語群
Table 1　Sentiment scales and their original sentiment words.

センチメント尺度 e	感情語群 e_1 ⇔ e_2
a：明るい	明るい，嬉しい，楽しい
⇔ 暗い	⇔ 暗い，悲しい，苦しい
b：承認	承認（する），愛好（する），好きだ
⇔ 拒否	⇔ 拒否（する），嫌悪（する），嫌いだ
c：緩和	ゆったり（する），のんびり（する），ゆっくり（する）
⇔ 緊張	⇔ 緊張（する），緊急（だ）
d：怒り	怒る，怒号
⇔ 恐れ	⇔ 恐れる，怖い，恐怖

表 2　センチメント辞書のエントリの例
Table 2　Example of sentiment dictionary entries.

	a：明るい⇔暗い		b：承認⇔拒否		c：緩和⇔緊張		d：怒り⇔恐れ	
内容語 w	$S_a(w)$	$M_a(w)$	$S_b(w)$	$M_b(w)$	$S_c(w)$	$M_c(w)$	$S_d(w)$	$M_d(w)$
蘇生（サ変名詞）	0.91	0.464	0.521	0.582	0.429	0.732	0.000	0.328
出国（サ変名詞）	0.596	0.975	0.209	1.049	0.762	1.065	0.201	0.701
死亡（サ変名詞）	0.28	1.132	0.358	1.272	0.260	1.306	0.364	1.112
脱線（サ変名詞）	0.31	0.514	0.546	0.603	0.403	0.737	0.291	0.549
出かける（動詞）	0.639	1.430	0.754	1.394	0.887	1.304	0.590	1.114
挑戦する（動詞）	0.618	1.399	0.687	1.330	0.752	1.251	0.500	1.090
衝突する（動詞）	0.344	1.004	0.353	1.016	0.315	1.099	0.529	0.948
懸念する（動詞）	0.373	1.447	0.319	1.440	0.246	1.521	0.293	1.275
豊富だ（形容詞）	0.597	1.416	0.676	1.352	0.761	1.299	0.466	1.109
最適だ（形容詞）	0.622	1.185	0.671	1.164	0.743	1.145	0.192	0.899
困難だ（形容詞）	0.318	1.451	0.305	1.526	0.307	1.528	0.317	1.274
不明だ（形容詞）	0.359	1.241	0.367	1.337	0.336	1.364	0.359	1.18

とを示す．以下，センチメント値 $S_e(w)$ と重み $M_e(w)$ の算出方法について述べる．

y 年版に掲載された記事のうち，感情語群 e に含まれる感情語のいずれかを含む記事の数を $df(y,e)$，感情語群 e に含まれる感情語とセンチメント辞書において見出し語となる内容語 w の両方を含む記事の数を $df(y,e\&w)$ とすると，感情語群 e のいずれかが現れたときに内容語 w も現れる確率 $P(y,e\&w)$ は，

$$P(y,e\&w)=\frac{df(y,e\&w)}{df(y,e)}$$

と表される．そこで，センチメント尺度 e を構成する感情語群 e_1，e_2 に対し，内容語 w の感情語群 e_1 に対する出現確率 $P(y,e_1\&w)$ と感情語群 e_2 に対する出現確率 $P(y,e_2\&w)$ の内分比 $R_{e_1\Leftrightarrow e_2}(y,w)$ を

$$R_{e_1\Leftrightarrow e_2}(y,w)=\frac{P(y,e_1\&w)}{P(y,e_1\&w)+P(y,e_2\&w)}$$

という式で求める．ただし，分母 $=0$ のときは，便宜的に $R_{e_1\Leftrightarrow e_2}(y,w)=0$ として処理する．

この $R_{e_1\Leftrightarrow e_2}(y,w)$ 値を年版ごとに求め，以下の式に代入することにより，内容語 w のセンチメント尺度 e におけるセンチメント値 $S_e(w)$ が求められる．

$$S_e(w)=\sum_{y=1990}^{2001}R_{e_1\Leftrightarrow e_2}(y,w)\Bigg/\sum_{y=1990}^{2001}T_{e_1\Leftrightarrow e_2}(y,w)$$

ただし，$T_{e_1\Leftrightarrow e_2}(y,w)$ は，$df(y,e_1\&w)+df(y,e_2\&w)=0$ のとき，0，そうでないとき，1 となる関数である．出現する場合には特定の感情語との結び付きが強いものも見受けられることから，導入されている．たとえば，特定の年度のみに出現するオリンピック関連用語に対して，$\sum_{y=1990}^{2001}T_{e_1\Leftrightarrow e_2}(y,w)$ はより小さい値をとるため，$S_e(w)$ は大きくなる．センチメント値 $S_e(w)$ は 0〜1 の値をとり，感情語群 e_1 と多くの記事に同時に出現した単語のセンチメント値が 1 に近くなり，感情語群 e_2 と多くの記事に出現した単語のセンチメント値が 0 に近くなる．

一方，内容語 w の中には，出現する年や出現頻度が多いものもあり，少ないものもある．そこで，センチメント値 $S_e(w)$ に対する重み $M_e(w)$ を以下のように定義し，内容語 w と感情語群 e_1，e_2 とが共起した年数と頻度の総和に応じて，増減するように設計した．

$$M_e(w)=log_{12}\sum_{y=1990}^{2001}T_{e_1\Leftrightarrow e_2}(y,w)\times log_{144}\sum_{y=1990}^{2001}(df(y,e_1\&w)+df(y,e_2\&w))$$

出現する年数と出現頻度が多いほど，重み $M_e(w)$ は高くなる．また，重み $M_e(w)$ が 0 である単語はセンチメント辞書に登録しない．

3.2　ニュース記事のセンチメント値の算出

ニュース記事 P のセンチメントは，($O_a(P)$, $O_b(P)$, $O_c(P)$, $O_d(P)$) といった形式を持つ．P が与えられると，日本語形態素解析システムを用いて，P に含まれるサ変名詞，動詞，形容詞を抽出する．P を抽出された単語 w の集合と見なし，各センチメント尺度 e（$e\in\{a,b,c,d\}$）に対してセンチメント辞書から各単語のセンチメント値 $S_e(w)$ と重み $M_e(w)$ を取得し，以下の式を用いて記事のセンチメント値 $O_e(P)$ を算出する．

$$O_e(P)=\sum_{w\in P}S_e(w)\times|2S_e(w)-1|\times M_e(w)\Bigg/\sum_{w\in P}|2S_e(w)-1|\times M_e(w)$$

この式は，$|2S_e(w)-1|\times M_e(w)$ を重みとするセンチメント値 $S_e(w)$ の重みつき平均であり，$|2S_e(w)-1|$ 項は $S_e(w)$ に依存する傾斜配分となっている．この傾斜配分は，感情語群との関係が乏しい一般的な単語（センチメント値は 0.5 に近い値をとる）が $O_e(P)$ 式の平均操作におよぼす悪影響を削減するために導入されている（センチメント値が 0.5 に近い値をとる単語の $|2S_e(w)-1|\times M_e(w)$ は 0 に近くなる）．

4. 検索処理

ユーザが検索キーワードを入力すると，システムは以下の処理を行い，センチメントマップを生成し，ユーザに提示する．

(1)　収集されたニュース記事からキーワードを含む記事をニュースサイトごとに検索する．

(2)　各ニュースサイトにおいて，検索キーワードの $tf\cdot idf$ 値を用いて記事をランキングする．

(3)　各ニュースサイトに対して，$tf\cdot idf$ 値の高い n 件のニュース記事を選び，3.2 節に述べた手法で算出されたセンチメント値を取得し，ニュースサイトごとに集計する．具体的には，1 日ごとに，各センチメント尺度に対して，サイト内のニュース記事のセンチメント値の平均値をとる．なお，ニュース記事のセンチメント値は 0〜1 の値をとるため，ニュースサイトのセンチメント値（サイト内のニュース記事のセンチメン

表 3　実験に用いたニュースサイト
Table 3　News websites used in the experiments.

国	地方	都道府県	新聞社名	URL
日本	北海道・東北	北海道	北海道新聞	http://www.hokkaido-np.co.jp/
		宮城	河北新報	http://www.kahoku.co.jp/
	関東・中部	東京	東京新聞	http://www.tokyo-np.co.jp/
		愛知	中日新聞	http://www.chunichi.co.jp/
	近畿・中国	兵庫	神戸新聞	http://www.kobe-np.co.jp/
		広島	中国新聞	http://www.chugoku-np.co.jp/
	九州	長崎	長崎新聞	http://www.nagasaki-np.co.jp/
		沖縄	沖縄タイムス	http://www.okinawatimes.co.jp/
中国			人民網（日本語版）	http://j1.people.com.cn/
韓国			朝鮮日報（日本語版）	http://www.chosunonline.com/
アメリカ			U. S. FrontLine（日本語版）	http://www.usfl.com/

ト値の平均値）は 0〜1 の値となる．また，次のステップで作成するセンチメントグラフの見やすさと対称性を考慮し，ニュースサイトのセンチメント値を −5〜+5 の値に調整した（元のニュースサイトのセンチメント値引く 0.5 掛ける 10）．

(4) JpGraph[9] を用いて，各ニュースサイトのセンチメントグラフを作成し，Yahoo Map API[10] を用いて，ニュースサイトの所在地により地図上にマッピングする．

(5) ユーザが地図をズームイン・ズームアウトすると，表示された範囲によって，ドメインを再決定し，センチメント値を再計算する．

本論文では，ニュース記事を対象にしており，ドメインの最小をニュースサイトとし，ドメインの最大を国とする．地図のズームイン・ズームアウトは世界地図，日本地図および各県地図といった 3 段階ある．世界地図が表示されるレベルでは，国ごとのグラフを表示する．日本地図が表示されるレベルでは，北海道・東北，関東・中部，近畿・中国，九州といった 4 つの地方のグラフを表示する．さらに，都道府県が表示されるレベルでは，東京新聞，神戸新聞といった各ニュースサイトのグラフを表示する．

各グラフは，最小の単位となるニュースサイトに対するセンチメント値を利用して，生成される．県レベルでの表示となる新聞社のグラフは，ニュースサイトのグラフをそのまま利用する．地方レベルでの表示のグラフは，各地方に含まれるニュースサイトのセンチメント値の平均値とし，グラフを生成する．国レベルでの表示のグラフは，各地方のすべてのニュースサイトのセンチメント値の平均値とし，グラフを生成する．

5. システムの実装と評価実験

特定トピックに対するニュース記事のセンチメント値を抽出し，地図の詳細度制御に基づく観点の相違を可視化するセンチメントマップシステムのプロトタイプを構築した．記事収集の対象としたニュースサイトと所在地の対応関係を**表 3** に示す．以下，システムのインタフェース，センチメント抽出の精度，地域ごとのセンチメント差異の評価について述べる．

5.1 センチメントマップのインタフェース

構築したシステムのインタフェースを**図 5** に示す．検索用インタフェースでは，収集対象のニュースサイトとニュース記事の取得日を提示する．これにより，ユーザは検索キーワードの入力（複数のキーワードを用いた「AND」検索も可能）と，ニュース記事の検索期間を選択できる．

図 6 は検索キーワードを「中国」に，検索期間を「2008 年 9 月 8 日〜9 月 10 日」にした場合の結果である．右上側フレームには，各ニュースサイトごとに，検索キーワードを含む $tf \cdot idf$ 値の高い 10 件の記事を提示した．右下側フレームには，各記事のセンチメントグラフとその記事に対して $tf \cdot idf$ 値の高い単語 10 個を提示した．左側フレームには，検索キーワードに対する各地域の 1 日ごとのセンチメントグラフを提示した．各センチメントグラフは，4 つのセンチメントの平均値とした．各グラフの上にマウスを置くと，4 次元のセンチメントとなる詳細なグラフを閲覧できる．また，ユーザがズームインやズームアウトすると，ニュースサイトレベルや国レベルのセンチメントマップが提示される．

図 5　プロトタイプのインタフェース
Fig. 5　Prototype interface.

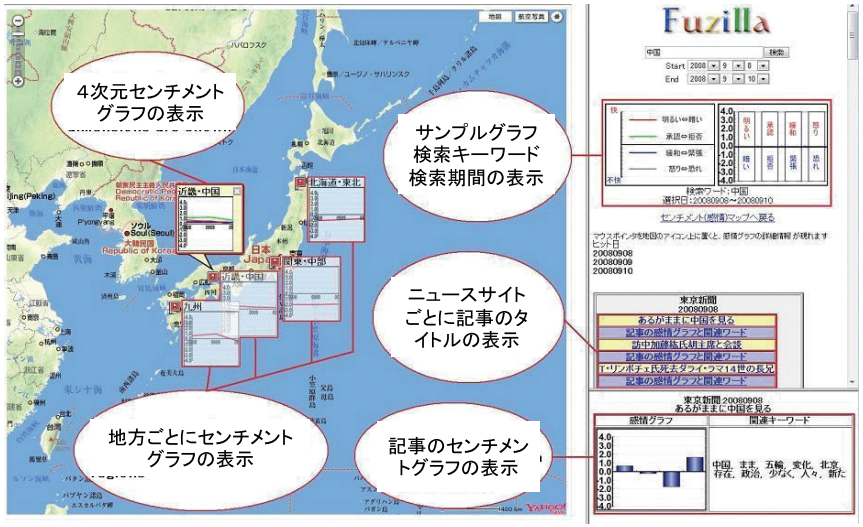

図 6　検索結果
Fig. 6　Retrieval result.

5.2　センチメント抽出精度の評価

システムのセンチメント抽出の精度を評価するため，各検索キーワードの記事に対してシステムが算出したセンチメント値と 100 名の一般ユーザ（20 代〜60 代の男女）の評価値との誤差を評価した．

システムが算出したセンチメント値は，検索キーワードの $tf \cdot idf$ 値が高い上位 10 件の記事を選出し，4 つのセンチメント尺度ごとに記事のセンチメント値を算出し，各尺度ごとのそれら 10 個のセンチメント値の平均値とした．

ユーザの評価値は，ユーザに各記事を閲覧してもらい，その記事の 4 つの尺度に対して 5 段階評価を行ってもらった．5 段階評価は，たとえば「明るい⇔暗い」の尺度の場合，「明るい」，「どちらかというと明るい」，「どちらでもない」，「どちらかというと暗い」，「暗い」とした．また，システムの算出値は 0〜1 の値を持つため，ユーザの 5 段階評価を同様に 0〜1 とするため，0，0.25，0.5，0.75，1，とした．具体的には，「明るい⇔暗い」の評価の場合，「明るい」= 1，「どちらかというと明るい」= 0.75，「どちらでもない」= 0.5，「どちらかというと暗い」= 0.25，「暗い」= 0 となる．次に，1 つの記事に対して，各尺度ごとに 100 名の評価値を集計した．算出は，5 段階評価の各々の人数を n_1，n_2，n_3，n_4 と n_5（$\sum_{i=1}^{5} n_i = 100$）とし，任意の記事に対するユーザの評価値を $(n_1*1+n_2*0.75+n_3*0.5+n_4*0.25+n_5*0)/100$ とした．最後に任意の検索キーワードに対する 10 件の記事の評価値の平均値をとり，各キーワードに対するセンチメント値をユーザの評価値とした．

表 4 にシステムの算出値とユーザの評価値との比較結果を示す．表 4 より，各センチメント尺度に対して，ユーザの評価値とシステムの計算値が近いことが確認できる．たとえば，キーワード「北京」，センチメント尺度「明るい⇔暗い」に対して，システムの算出値（0.5110）はユーザの評価値（0.5203）と近い数値となっている．各センチメント尺度に対する平均誤差は 7%〜10%となり，被験者のセンチメントと近い値を算出できていることが確認でき，高い精度でセンチメント値を抽出できたことを示せた．ただし，「怒り⇔恐れ」の尺度における平均誤差は 10%と高めであり，精度の改善が必要と考えられる．また，トピック単位では，「教員」のキーワードに関してユーザ評価値とシステム算出値との間にずれが生じている．評価実験で被験者に提示した記事は「教員採用汚職」に関するものであっ

表 4　システムが与えたセンチメント値とユーザの評価値との誤差
Table 4　Error of sentiment values given from the system and individuals.

検索キーワード		検索キーワードに対するセンチメント値			
		明るい⇔暗い	承認⇔拒否	緩和⇔緊張	怒り⇔恐れ
北京	ユーザ評価値	0.5203	0.5815	0.5368	0.5165
	システム算出値	0.5110	0.5255	0.3766	0.5566
教員	ユーザ評価値	0.2533	0.3230	0.3528	0.7560
	システム算出値	0.4639	0.5135	0.4430	0.4952
橋下知事	ユーザ評価値	0.5080	0.5590	0.5115	0.5075
	システム算出値	0.4236	0.5244	0.5238	0.4571
京都	ユーザ評価値	0.4733	0.5903	0.5135	0.5120
	システム算出値	0.5299	0.5440	0.3983	0.5587
福田首相	ユーザ評価値	0.4418	0.4825	0.4453	0.5800
	システム算出値	0.4208	0.4692	0.4957	0.4519
平均誤差		0.07638	0.06814	0.08566	0.10522

たため，ユーザ評価値は極端にネガティブなセンチメント値をとっているが，システムによるセンチメント値の算出では，記事全体の単語に着目した結果，「汚職」に類するネガティブな単語だけでなく，「教員」や「採用」といったポジティブな単語も含まれていたため，あまりネガティブなセンチメントに寄らなかったものと考えられる．これらの原因の 1 つとして，単語ベースでセンチメント値を計算していることが考えられることから，現在，句レベル（動詞句，形容詞句，名詞句）のセンチメント値算出手法を検討している．

5.3　地域ごとのセンチメント差異の評価

前節と同様に，100 名のユーザが地図の詳細度制御に対する発信者の観点の理解に関して評価実験を行った．評価方法は，地図に提示されているセンチメントマップを閲覧してもらい，3 つのトピックに対して，センチメントの相違が理解できたか否かを 5 段階評価してもらった．**図 7** に評価結果を示す．各々の観点の理解に関して，「できた」「どちらかといえばできた」の割合は 40%〜50%であった．「まったくできていない」「どちらかといえばできていない」の割合が 25%〜35%であった．これより，ユーザが各々のセンチメントの相違を理解していることが確認できた．

上記の定量的な評価のほか，提案システムが提示するニュースサイトごとのセンチメント差異に関して，著者らにより定性的な評価を行った．トピックとして**表 5** に示す 20 トピックを使用し，システムが提示した結果に対して，ニュースサイト間のセンチメント差異が存在したか否かを評価した．20 トピックのうち，4 トピック（「羽田ハブ化」，「日本シリーズ」，「内藤大助」，「松井秀」）に関しては，ニュースサイト間での明らかなセンチメント差

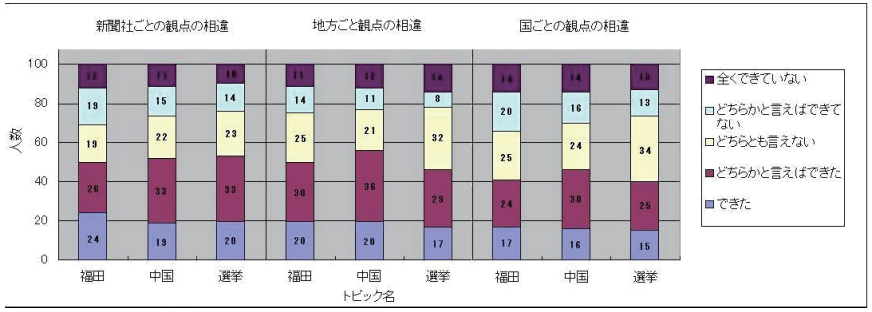

図 7　地域ごとのセンチメント差異の評価
Fig. 7　Evaluation of sentiment distinction between different regions.

表 5　考察で使用された 20 トピック
Table 5　20 topics used in the experiment.

羽田ハブ化	日本シリーズ	内藤大助	松井秀
事業仕分け	ダム中止	高速道路無料化	たばこ税
日航支援	新型インフルエンザ	温暖化対策	偽装献金
市橋容疑者	エコ減税	年末ジャンボ	景気回復
サッカー W 杯	中村俊輔	福原愛	石川遼

異が見られたが，ほかの 16 トピックに関して，大きなセンチメント差異はなかった．多くのトピックに対して，新聞社が事実を述べ，他社と同様な傾向で記事を書くことが多いと考えられる中で，ニュースサイト間の差異を発見でき，本システムの有用性が示せた．

センチメント差異があった 4 トピックのセンチメント値を**表 6** に示す．なお，表 6 に示したセンチメント値は，4 章のステップ (3) に述べた方法で算出された −5〜+5 の調整値である．他社と比べて，相対的に大きい・小さいセンチメント値をとったニュースサイトのセンチメント値を太字で表した．センチメント差異が見られた原因として，次のようなものが考えられる．

- 「羽田ハブ化」に対して，東京新聞は「承認⇔拒否」の尺度において，他社より大きいセンチメント値をとった．地元新聞が支持する偏向を把握できるとともに，異なる傾向のサイトの存在も確認でき，立場の違いを把握しつつ多様な観点の記事を閲覧できる．
- 「日本シリーズ」に対して，北海道新聞は「明るい⇔暗い」の尺度において，他社より小さいセンチメント値をとった．これは，読売ジャイアンツの勝利を多く取り上げる他

表 6　ニュースサイト間のセンチメント差異を発見できた例
Table 6　Examples with sentiment distinction between news websites.

トピック	センチメント尺度	東京新聞	神戸新聞	北海道新聞	河北新報	中日新聞	U. S. FrontLine
羽田ハブ化	承認⇔拒否	**0.21518**	−0.05207	0.073	0.01712	0.00509	−0.04169
日本シリーズ	明るい⇔暗い	2.00288	2.06228	**1.65161**	2.11594	1.97051	No data
内藤大助	明るい⇔暗い	1.12832	0.7919	**1.26834**	1.14101	1.05233	No data
松井秀	明るい⇔暗い	1.9682	1.63669	2.0173	1.65734	1.88283	**1.59814**

社と比べて，北海道新聞は日本ハムファイターズの立場から，敗戦を報道する記事が少なからずあったためと考えられる．

- 「内藤大助」に対して，北海道新聞社のセンチメントは他社より明るかった．これは，内藤選手の出身地である北海道新聞社が内藤選手に対する応援や激励を記事として取り上げることが多かったためと考えられる．
- 「松井秀」に対して，アメリカのニュースサイトのセンチメントは日本の新聞社ほど明るくなかったことが分かる．これは，日本の新聞社は松井選手の活躍を中心に報道するのに対して，アメリカのニュースサイトはより多面的に報道するためと考えられる．

その他の 16 個のトピックに対しては，ニュースサイト間のセンチメント差異が見られなかったが，少数でもそのような差異があるトピックを発見できることが重要であり，提案システムの有効性を示せた．

6. 関 連 研 究

近年，多くのニュースポータルサイト[4),5)]が増えてきた．これらのサイトは，ニュース記事の収集，分類，統合や推薦などを行っている．Google ニュース[4)]は約 4500 個のニュースサイトから記事を収集し，類似する記事を提供している．Yahoo!ニュース[5)]は読者のコメント，ブロガの注目やブックマークの数などの情報に基づきニュース記事のランキングを行っている．ユーザはこれらのポータルサイトを利用することにより，興味があるニュース記事を閲覧できる．しかし，これらのポータルサイトはニュース記事の発信者の観点の差異を考慮していない．

一方では，センチメント分析に関する研究[11)–13)]がある．センチメント分析技術は，映画レビューや商品評価などのテキストデータからセンチメントを抽出する．Turney は，各種レビューを「recommended」か「not recommended」に分類する手法[14)]を提案している．彼の手法は，入力テキストから特定パターン（たとえば「形容詞＋名詞」や「副詞＋形容詞＋名詞以外」など）のフレーズを抽出し，各フレーズと参照語「excellent」および「poor」との相互情報量をそれぞれ求め，差をとることにより，各フレーズの Semantic Orientation（SO）を決定している．そして，全フレーズの SO を平均することにより，入力テキストの SO を求め，「recommended」か「not recommended」かを決定している．彼の手法で特徴的なのは，レビューの SO を特定の 2 単語「excellent，poor」との対応関係を調べることにより算出したという点と，その対応関係を，辞書の類を参照するのではなく，AltaVista Advanced Search engine におけるヒット数を用いて調べたという点にある．Pang ら[15)]は映画レビューの文書から主観的な部分のみを抽出し，テキスト分類の技術を用いて「好評」か「不評」に分類した．Esuli ら[16)]はオンライン辞書を用いて単語の極性（ポジティブまたはネガティブ）を判断した．しかし，これらの研究は文書や単語の肯定・否定分類に限られている．本研究は，より人間の感情に近いとされる 4 次元のセンチメントを分析する．また，そのセンチメントの差異を地域性に基づき可視化する．

本研究が用いた Plutchik の感情モデル[8)]のほかに，様々な感情モデルが提案されている[17),18)]．Russell[17)]は横軸を「愉快—不快」，縦軸を「興奮—眠気」とした 2 次元の感情空間を提案した．「刺激」，「消沈」，「安らぎ」，「嘆き」といった 4 つの感情は独立な軸ではなく，「愉快」，「不快」，「興奮」，「眠気」の組合せにより生成されたものとした．Pitel ら[18)]は 44 個の感情ペアを考慮し，SVM 分類器を用いてフランス語のセンチメント辞書を構築した．これらのモデルとの比較は今後の課題の 1 つとなる．

7. まとめと今後の課題

本論文では，ニュース記事を対象にセンチメントを抽出し，任意のトピックに関して抽出したセンチメントを各サイトごとに分析することで，発信者の観点差異を発見・可視化するセンチメントマップシステムを提案した．実験では，センチメントマップにより，発信者のセンチメントの相違を発見できることを確認できた．発信者の観点差異を発見・提示するこ

とは，ニュース記事の記述の公平性や信頼性を判断する材料になると考えられる．

システムの現状では，文書単位で記事のセンチメント値を求めたが，今後の課題として，段落単位や文単位で分析することで，センチメント値を算出する精度を向上させたい．また，ニュースサイトのセンチメント傾向を提供するために，サイト内の記事のセンチメントの平均値を求めたが，サイト内のばらつきを表現できていない．今後は，サイト内のセンチメントの標準偏差も提示するなど，より詳細な分析・提示を行いたい．

謝辞　本研究の一部は，独立行政法人情報通信研究機構による委託研究「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」の成果であり，ここに記して謝意を表す．

参 考 文 献

1) 読売新聞．http://www.yomiuri.co.jp/
2) 朝日新聞．http://www.asahi.com/
3) 毎日新聞．http://mainichi.jp/
4) Google ニュース．http://news.google.co.jp/
5) Yahoo!ニュース．http://headlines.yahoo.co.jp/
6) 河合由起子，官上大輔，田中克己：個人の選好に基づく複数ニュースサイトの記事収集・閲覧システム，情報処理学会論文誌：データベース，Vol.46, No.SIG8 (TOD26), pp.14–25 (2005).
7) 河合由起子，熊本忠彦，田中克己：印象と興味に基づくユーザ選好のモデル化手法の提案とニュースサイトへの応用，知能と情報（日本知能情報ファジィ学会誌），Vol.18, No.2, pp.173–183 (2006).
8) Plutchik, R.: *The Emotions*, Univ. Pr. of Amer. (1991).
9) JpGraph. http://www.asial.co.jp/jpgraph/
10) Yahoo! Map API. http://developer.yahoo.co.jp/webapi/map/
11) Strapparava, C. and Mihalcea, R.: Task 14: Affective Text, *SemEval* (2007).
12) Pang, B. and Lee, L.: Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval*, Vol.2, No.1-2, pp.1–135 (2007).
13) Wright, A.: Our Sentiments, Exactly, *Comm. ACM*, Vol.52, No.4, pp.14–15 (2009).
14) Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *ACL*, pp.417–424 (2002).
15) Pang, B. and Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *ACL*, pp.271–278 (2004).
16) Esuli, A. and Sebastiani, F.: Determining the Semantic Orientation of Terms through Gloss Classification, *CIKM*, pp.617–624 (2005).
17) Russell, J.A.: A Circumplex Model of Affect, *Journal of Personality and Social Psychology*, Vol.39, No.6, pp.1161–1178 (1980).
18) Pitel, G. and Grefenstette, G.: Semi-automatic Building Method for a Multidimensional Affect Dictionary for a New Language, *LREC* (2008).

(平成 21 年 9 月 20 日受付)
(平成 21 年 12 月 30 日採録)

(担当編集委員　関 洋平)

張　　建偉（正会員）
京都産業大学コンピュータ理工学部特定研究員．2005 年筑波大学大学院システム情報工学研究科博士前期課程修了．2008 年筑波大学大学院システム情報工学研究科博士後期課程修了．博士（工学）．ウェブマイニング，信憑性分析の研究に従事．日本データベース学会会員．

河合由起子（正会員）
京都産業大学コンピュータ理工学部講師．2001 年奈良先端科学技術大学院大学情報科学研究科情報システム学博士後期課程修了．同年独立行政法人情報通信研究機構，2006 年京都産業大学理学部コンピュータ科学科講師を経て 2008 年より現職．博士（工学）．情報推薦，Web マイニング，信憑性分析の研究に従事．日本データベース学会会員．

熊本　忠彦（正会員）
1988 年筑波大学第三学群情報学類卒業．1990 年筑波大学大学院理工学研究科修士課程修了．同年郵政省通信総合研究所（現，独立行政法人情報通信研究機構）入所．2007 年千葉工業大学情報科学部情報ネットワーク学科准教授．現在に至る．近年は，印象マイニングとその応用（情報検索，記事推薦，アニメーション生成，ユーザモデリング，ほか）に関する研究に従事．1996 年博士（工学）（筑波大学）．FIT2004 論文賞受賞．電子情報通信学会，人工知能学会，言語処理学会，日本データベース学会各会員．

田中　克己（正会員）

京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了，1978 年同博士後期課程中退．1981 年京都大学工学博士．主にデータベース，Web 情報検索，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，日本データベース学会各会員．