Vol.2012-NL-207 No.4
2012/7/26

ユーザ印象評価データの観察と分析に基づくテキスト印象マイニング手法の設計

熊本 忠彦[1]　河合 由起子[2]　張 建偉[3]

概要：本稿では，「楽しい ⇔ 悲しい」，「うれしい ⇔ 怒り」，「のどか ⇔ 緊迫」という 3 種類の印象を対象に，新聞記事を読んだ人々が感じる印象の強さを数値的に求めるための印象マイニング手法を提案する．提案手法は，各々の印象に対して「(左側の印象を) 感じる (1 点)，割と感じる (2 点)，やや感じる (3 点)，(どちらの印象も) 感じない (4 点)，(右側の印象を) やや感じる (5 点)，割と感じる (6 点)，感じる (7 点)」という 7 段階の評価スケール（本稿では，印象尺度と呼ぶ）に準じた 1.0〜7.0 の実数値を出力する．このような手法の設計に際し，記事から抽出される特徴量（単語 unigram）とあらかじめ定義した特定の印象語群との（記事内）共起関係に基づいて各印象尺度用の印象辞書（各単語 unigram の記事印象への影響力を示すもの）を構築し，それぞれの印象辞書を用いて算出される記事の印象値とその記事を読んだ人々が感じる印象の強さとの対応関係を定式化することで，共起関係という読み手が介在しない方法で算出される記事の印象値を読み手が感じる印象の強さへと変換することを試みる．本手法の未知データに対する誤差（RMSE：Root-Mean-Square Error）を 5 分割交差検定により調べてみたところ，それぞれの印象尺度に対し 0.60，0.49，0.52 であった．従来手法の誤差は 0.69，0.49，0.64 であり，「うれしい ⇔ 怒り」に対しては同じ誤差を保ちつつ，「楽しい ⇔ 悲しい」と「のどか ⇔ 緊迫」に対する誤差が大幅に改善されているのがわかる．

1. はじめに

近年，テキストから評判や感情，印象といった主観的な情報を抽出するための研究が盛んであり，評判分析 [1], [2] や情報可視化 [3]，印象アノテーション [4], [5] などに応用されている．しかしながら，ほとんどの研究がクラス分類問題あるいはアノテーション問題としての扱いであり，読み手が感じる印象（特に「悲しい」や「のどか」などの感情的印象）の強さを数値的に求めるという研究はまだ少ない [6], [7]．印象の強さを数値化することにより，印象空間（複数の印象尺度によって張られる多次元空間）へのテキストデータの写像が可能となり，その結果，テキストデータに対する印象分布の可視化やランキングといった操作が可能となる．

そこで本稿では，新聞記事を例に，記事を読んだ人々が感じる印象の強さを数値的に求めるための印象マイニング手法を提案する．本手法が対象とする印象は，「楽しい ⇔ 悲しい」，「うれしい ⇔ 怒り」，「のどか ⇔ 緊迫」の 3 種類であり，それぞれの印象に対して「(左側の印象を) 感じる (1 点)，割と感じる (2 点)，やや感じる (3 点)，(どちらの印象も) 感じない (4 点)，(右側の印象を) やや感じる (5 点)，割と感じる (6 点)，感じる (7 点)」という 7 段階の評価スケール（以下，印象尺度と呼ぶ）を割り当てる．すなわち，提案手法は，それぞれの印象尺度において，そのスケールに準じた 1.0〜7.0 の実数値を，印象の強さを表す印象値として出力する．例えば，提案手法の「のどか ⇔ 緊迫」に対する出力値が 2.30 であった場合，その記事ののどかさに関しては，「割と感じる (2 点)」よりやや「やや感じる (3 点)」寄りであると判断されたことがわかる．

このような手法を設計するに当たり，出力された印象値が，読み手が感じる印象の強さを表していることをどうやって担保するのかという問題を考える．著者らは，新聞記事データベースから抽出される任意の特徴量とあらかじめ定義した特定の印象語群との（記事内）共起関係に基づいて印象辞書 [8]（記事から抽出される特徴量の記事印象への影響力を示すもの）を印象尺度ごとに構築し，それぞれの印象辞書を用いて算出される記事の印象値とその記事を読んだ人々が感じる印象の強さとの対応関係を定式化することで，共起関係という読み手が介在しない方法で算出さ

[1] 千葉工業大学
Chiba Institute of Technology
[2] 京都産業大学
Kyoto Sangyo University
[3] 筑波技術大学
Tsukuba University of Technology

れる記事の印象値を読み手が感じる印象の強さへと変換することを試みる．具体的には，以下の 2 つのアプローチを採用する．

まず，最初のアプローチとして，それぞれの印象尺度において印象辞書を用いて算出される記事の印象値とその記事を人々が読んだときに感じる印象の強さとの対応関係を回帰分析により調べ，その結果得られる回帰式を用いて算出された印象値を補正する，ということを考える．このアプローチは，著者らの先行研究 [8] ですでに提案しており，その有効性も確認されている．次に，2 つ目のアプローチとして，それぞれの印象尺度において算出され，補正された印象値を相互に利用し，各印象尺度における記事の印象値を算出し直す，ということを考える．本稿では，この 2 つのアプローチを組み合わせることで，より高精度な手法の実現を目指す．なお，本稿では，記事から抽出する特徴量として単語 unigram を採用する．単語 unigram は，記事特徴量としての網羅性が高く，使い勝手がいいことから，数多くの研究 [2], [8] で採用されている．特に Pang らの研究 [2] では他の特徴量（単語 bigram のみ，単語 unigram と単語 bigram の組み合わせ，単語 unigram と品詞情報との組み合わせなど）を用いるよりも単語 unigram のみを用いた方が，精度が良かったことが示されている．

本稿の構成を示す．まず，第 2 章で関連研究を整理し，提案手法との違いを示す．第 3 章で 900 人が参加するアンケート調査を行い，印象評価データ（全 90 記事の印象の強さを数値化したもの）を作成する．第 4 章で新聞記事データベースから印象辞書を構築する手法，印象辞書を用いて記事の印象値を算出する手法，算出された記事の印象値を回帰式を用いて補正する手法 [8] について述べる．第 5 章で各印象尺度において算出され，補正された記事の印象値を説明変数（3 個），アンケート調査に基づいて作成された印象評価データを目的変数とする重回帰分析を行い，その対応関係を定式化する．第 6 章で提案手法の精度を評価し，従来手法 [8] の精度と比較することで，その有効性を検証する．最後に，第 7 章で本稿のまとめと今後の課題について述べる．

2. 関連研究

映画レビューや書評といった書き手の評価を Positive, Negative の 2 クラス，あるいは Neutral を加えた 3 クラスに分類するという研究がある．例えば，Turney[1] は，各種レビューを「recommended」か「not recommended」に分類する手法を提案している．彼の手法は，入力テキストから特定パターン（例えば「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」との自己相互情報量 [9] をそれぞれ求め，差を取ることにより，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，その値により「recommended」か「not recommended」かを決定している．しかしながら，この手法は，印象の強さを数値化するのではなく，クラスへの分類問題として扱っている点や 1 本の印象尺度（「recommended ⇔ not recommended」）に特化し，複数の印象を対象としていない点が異なっている．

一方，テキストを複数のクラスに分類するという研究も行われている．例えば，Lin ら [3] は，ニュース記事を 8 つの感情クラス（Awesome，Heartwarming，Surprising，Sad，Useful，Happy，Boring，Angry）に分類する SVM（Support Vector Machine）ベースの手法を提案している．具体的には，提示された中国語のニュース記事に対して，指定された 8 つの感情のうちの 1 つを投票できる Web 上のニュースサイトを利用することで，それぞれのニュース記事本来の感情を決定し，各ニュース記事から抽出される特徴量（単語 unigram，文字 bigram，affix similarity など）と関連付けたものを SVM への訓練データとして用いている．つまり，読み手を意識した研究であり，複数の感情を対象としている点で，著者らの研究と同じ方向性を持っているが，SVM によるクラス分類問題として扱っている点が印象の強さの数値化を目指している著者らの研究とは異なっている．

入力されたテキストに任意の印象語をタグとして付与するという研究がある．例えば，宮川ら [4] は，意味の数学モデル [10] を用いてテキストが有する任意の印象を抽出するための手法を提案している．この意味の数学モデルは，文脈に応じた意味的連想を可能とする情報検索方式であり，検索に用いられるキーワード群を配置したメタデータ空間と呼ばれる正規直交空間から文脈を表す部分空間を選択し，その部分空間上での相関量に基づいて意味的に近いキーワード（印象語）の選択を可能にしている．この方式では，テキストの印象を表すキーワード（印象語）を文脈に応じて選択することが可能と考えられるが，特定の印象尺度に沿って印象の強さを数値的に求めることはできない．また，メタデータ空間を構築する際に，専門用語辞典や国語辞書といった言語資源における単語間の共起関係を用いているため，読み手を意識した手法とは言えない．一方，清水ら [5] は，特定のフレーズパターンの出現頻度に基づいて形容詞どうしの意味的関係や形容詞と名詞，形容詞と動詞との意味的関係を抽出し，さらに名詞と動詞の組み合わせに対する印象（特に「嬉しい」，「明るい」，「寒い」，「冷たい」，「重い」のような情景を表す形容詞）を推定する手法を提案している．この手法の特徴として，印象推定の信頼性を印象適合値という数値で表している点が挙げられるが，印象の強さを数値的に求めることはできない．また，言語資源として Web ページを用いているため，読み手を意識した設計とは言えない．

表 1 Juman の出力結果を変換するための後処理ルール [8]

1. 形容詞／動詞＋名詞性述語接尾辞のとき，この 2 語を普通名詞 1 語に変換する
2. 名詞／未定義語／形容詞（語幹）／動詞（基本連用形）＋名詞性名詞接尾辞（「化」を除く）のとき，この 2 語を普通名詞 1 語に変換する
3. 名詞／未定義語／形容詞（語幹）／動詞（基本連用形）＋名詞性名詞接尾辞「化」のとき，この 2 語をサ変名詞 1 語に変換する
4. 接頭辞「御／ご／お」＋動詞（基本連用形）のとき，この 2 語をサ変名詞 1 語に変換する
5. 動詞（基本連用形）＋格助詞のとき，動詞（基本連用形）をサ変名詞に変換する
6. 名詞／未定義語＋名詞性特殊接尾辞（「都，道，府，県，郡，市，町，村，区，州，省」を除く）のとき，この 2 語を未定義語 1 語に変換する
7. 形容詞／動詞＋動詞性接尾辞のとき，この 2 語を動詞 1 語に変換する
8. サ変名詞／カタカナ／アルファベット／副詞／形容詞（基本連用形／ダ列基本連用形）＋動詞「する／できる」のとき，この 2 語を動詞 1 語に変換する
9. 動詞＋助動詞「ぬ」／形容詞性述語接尾辞「ない」のとき，この 2 語を動詞 1 語に変換する
10. 名詞（形式名詞と副詞的名詞を除く）／未定義語／動詞（基本連用形）／副詞＋判定詞のとき，この 2 語を形容詞 1 語に変換する
11. 形容詞／動詞／判定詞＋形容詞性述語接尾辞（「ない」を除く）のとき，この 2 語を形容詞 1 語に変換する
12. 名詞／未定義語／動詞／形容詞＋形容詞性名詞接尾辞のとき，この 2 語を形容詞 1 語に変換する
13. 形容詞＋形容詞性述語接尾辞「ない」のとき，この 2 語を形容詞 1 語に変換する
14. 形式名詞／副詞的名詞／助詞＋判定詞のとき，この 2 語を判定詞 1 語に変換する
15. 判定詞＋形容詞性述語接尾辞「ない」のとき，この 2 語を判定詞 1 語に変換する
16. 形容詞（ダ列タ系連用テ形／基本連用形）／動詞（タ系連用テ形）／判定詞（ダ列タ系連用テ形）＋副助詞「は／も」のとき，副助詞を削除する
17. 接頭辞（「御，ご，お」を除く）＋任意の形態素のとき，この 2 語を 1 語にする

著者らの研究と同様，読み手がどう受け取るかという観点から読み手が感じる印象の強さを数値的に求めるための研究も行われている．例えば，阿部ら [6] は，5 種類の感情（喜，怒，哀，楽，愛）を対象に，受信したメールを読んだユーザが抱く感情の度合いを推定し，それぞれの感情の度合いに応じて楽曲を推薦する手法を提案している．この手法の感情推定部分を設計するに当たり，彼らは，訓練用のメールから抽出された特徴語（名詞，動詞，形容詞）のうち，tf・idf 値が閾値以上のものを説明変数，各メールに対し，被験者らが付けた各感情の評価点（5 段階評価尺度）を目的変数とする重回帰分析を感情の種類ごとに行い，それぞれの対応関係を重回帰式で表している．しかしながら，メールから抽出される特徴語をそのまま説明変数としているため，汎用性（新規受信メールへの対処能力）という点で問題が生じる．すなわち，重回帰分析では，説明変数の数より多い数の訓練事例が必要とされるが，この手法に汎用性を持たせるためには相当数の訓練事例が必要となり，実際的ではない．もし訓練事例数が十分でないと，その分，説明変数の数を絞り込む必要があり，その結果，新規の受信メールに説明変数として選ばれた特徴語があまり含まれず，正確に感情推定できないということも考えられる．これに対し，著者らの研究では，記事から抽出された特徴量を用いて，仮の記事印象値を算出し，この印象値を回帰分析のための説明変数として用いているので，記事から抽出する特徴量の数には制限がないという利点がある．一方，秋山ら [7] は，「かくかく」のような XYXY 型のオノマトペ（擬音語，擬態語，擬声語など）から感じる印象を 4 つの因子「キレ・俊敏さ」，「柔らかさ・丸み」，「躍動感」，「大きさ・安定感」で定義し，各因子を形容詞対からなる 5 段階評価尺度（例えば「躍動感のない ⇔ 躍動感のある」）で表すことで，それぞれの因子における印象の強さを数値的に求める手法を提案している．具体的には，14 種類の音の要素（子音 9 種類，母音 5 種類）と 38 種類の XYXY 型オノマトペに対する各因子の値を被験者実験で求め，各オノマトペに対する因子の値を目的変数，そのオノマトペを構成する文字 X と Y の子音と母音に対する因子の値を説明変数（計 4 個）とする重回帰分析を因子ごとに行うことで，音の要素とオノマトペから受ける印象の強さとの対応関係を定式化している．しかしながら，この手法は，オノマトペの音響的な特徴を利用した研究であり，かつ音の組み合わせ方（文献 [7] では XYXY 型のみが対象）に制限があることから，一般的なテキストへは応用できない．

3. ユーザ印象評価データの収集

人々が新聞記事からどのような印象を受けるのかを示す印象評価データを得るために，900 人（男女 450 人ずつ）が参加するアンケート調査を行った．具体的には，回答者 900 人を年齢や性別が均等になるよう 9 つのグループ（男女 50 人ずつ，計 100 人からなるグループ）に分け，各グループに毎日新聞の 2002 年版社会面 [11] に掲載された 10 記事を提示した．この 10 記事はグループによって異なっており，全部で 90 記事が重複しないように選ばれている．各回答者は，ランダムに提示される 10 記事の印象をランダムな順番で提示される 3 種類の印象尺度を用いて 7 段階

評価した．すなわち，「楽しい ⇔ 悲しい」，「うれしい ⇔ 怒り」，「のどか ⇔ 緊迫」のそれぞれに対し，対応する印象をどの程度感じるかを「(左側の印象を）感じる（1 点），割と感じる（2 点），やや感じる（3 点），(どちらの印象も）感じない（4 点），(右側の印象を）やや感じる（5 点），割と感じる（6 点），感じる（7 点）」の 7 段階で評価した．

以上の結果得られたデータから各記事の各印象尺度における平均値を求めた．本稿では，この平均値を記事本来の印象値とみなし，印象評価データとして扱う．

なお，各回答者に提示した記事は，元の記事の第 1 段落のみであり，第 2 段落以降は提示していない．これは，記事の構成上，第 1 段落を読めば記事の概要がわかるように書かれている点や段落ごとに記事の印象が変わる可能性がある点，記事が長いと回答者にかかる負担が増大する点を考慮した結果である．但し，将来的には 1 つの記事の中での印象の推移を追跡できるような印象マイニングを実現したいと考えている．

4. 記事印象値の算出と回帰式による補正

本章では，著者らの先行研究 [8] に基づいて，新聞記事から抽出する特徴量として単語 unigram を定義し，3 つの印象辞書（3 種類の印象尺度に対応）を構築するとともに，それぞれの印象尺度において，印象辞書を用いて算出される記事の印象値を説明変数，第 3 章のアンケート調査に基づいて数値化された記事本来の印象値を目的変数とする回帰分析を行い，両者の対応関係を表すのに最適な回帰式を得る．

4.1 単語 unigram の生成

はじめに，新聞記事から記事特徴量として単語 unigram を生成する手法について述べる．

まず，日本語汎用形態素解析システムである Juman [12] を用いて，入力された記事を単語に分解する．この Juman の出力結果に対し，表 1 に示したルールを再帰的に適用し，後処理を行う．例えば，「削除しない」のようなフレーズは，Juman によりサ変名詞「削除」，動詞「する」，形容詞性述語接尾辞「ない」の 3 語に分けられるが，ルール 8 とルール 9 を順に適用することにより，「削除しない」という動詞 1 語として扱うことになる．同様に，「楽しくはなかった」のようなフレーズは，形容詞「楽しい」，副助詞「は」，形容詞性述語接尾辞「ない」の 3 語に分けられるが，ルール 16 とルール 13 を順に適用することにより，「楽しくない」という形容詞 1 語として扱うことになる．また，「ホームランだ」のようなフレーズは，普通名詞「ホームラン」と判定詞「だ」の 2 語に分けられるが，ルール 10 を適用することにより，「ホームランだ」という形容詞 1 語として扱うことになる．以上の後処理の結果から助詞，連体詞，指示詞を取り除いたものが単語 unigram として利用される．なお，表 1 に示したルールは，著者らが直感的に定めたものであり，接尾辞や接頭辞が単語の印象に与える影響を考慮するためのものや品詞を変換するためのものなどがある．

表 2 各印象尺度を構成する印象語群 [8]

印象尺度	印象語群（上段：I_L，下段：I_R）
楽しい	楽しい，楽しむ，楽しみだ，楽しげだ
⇔ 悲しい	悲しい，悲しむ，悲しみだ，悲しげだ
うれしい	うれしい，喜ばしい，喜ぶ
⇔ 怒り	怒る，憤る，激怒する
のどか	のどかだ，和やかだ，素朴だ，安心だ
⇔ 緊迫	緊迫する，不気味だ，不安だ，恐れる

4.2 印象辞書の自動構築

次に，前節の方法で生成された単語 unigram を用いて印象辞書を自動構築する手法について述べる．

まず，「ある印象を有する単語 unigram は，その印象を表現する印象語群と共起しやすく，逆の印象を表現する印象語群とは共起しにくい」という仮定を置き，この仮定のもと，5 年分の読売新聞記事データ（2002 年版〜2006 年版）から生成される任意の単語 unigram U と印象尺度ごとに用意される対比的な印象の 2 つの印象語群との共起の仕方を調べ，どちらの印象語群とより共起しやすいかを数値化したものを，U の当該印象尺度における印象値として印象辞書に登録する．具体的な手順を以下に示す．

まず，各印象尺度の左側の印象（楽しい，うれしい，のどか）を表す印象語群 I_L と右側の印象（悲しい，怒り，緊迫）を表す印象語群 I_R を定義し，解析対象となる新聞記事データから印象語群 I_L あるいは I_R に含まれる印象語を 1 語以上含む記事を抽出するとともに，各記事に含まれる印象語の数を印象語群ごとに数える．

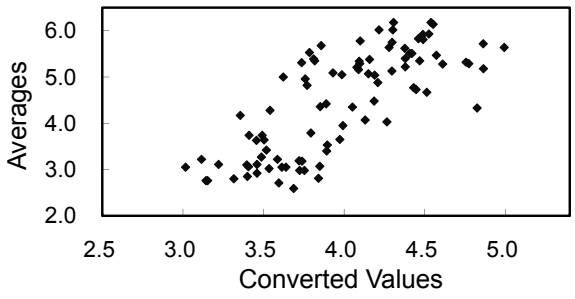
以上の結果，印象語群 I_L に属する印象語の数が印象語群 I_R に属する印象語の数よりも多かった記事の集合を S_L（記事数を N_L）とし，逆に少なかった記事の集合を S_R（記事数を N_R）とする．

次に，それぞれの記事集合（S_L もしくは S_R）からすべての単語を抽出し，前節で述べた手法を用いて単語 unigram を生成するとともに，その出現記事数を数える．このとき，ある単語 unigram U の記事集合 S_L における出現記事数を $N_L(U)$，記事集合 S_R における出現記事数を $N_R(U)$ とすると，それぞれの条件付き出現確率 $P_L(U)$ と $P_R(U)$ は次のように表される．

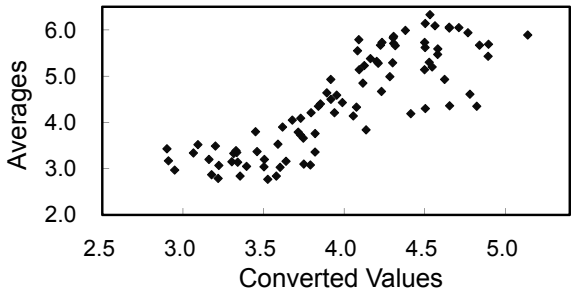
$$P_L(U) = \frac{N_L(U)}{N_L}$$

$$P_R(U) = \frac{N_R(U)}{N_R}$$

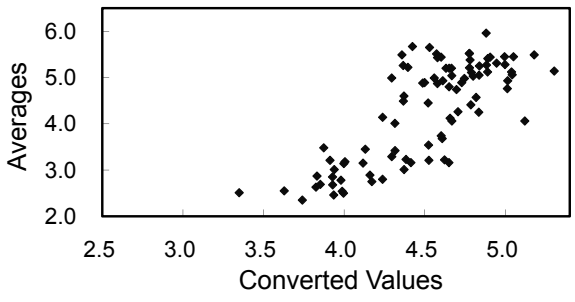
この $P_L(U)$ と $P_R(U)$ を用いて，単語 unigram U の印象値 $v(U)$ を以下の式で計算する．



(a)「楽しい ⇔ 悲しい」の場合



(b)「うれしい ⇔ 怒り」の場合



(c)「のどか ⇔ 緊迫」の場合

図 1 印象辞書を用いて算出された記事の印象値（換算値）と回答者らが付けた 7 段階評価値の平均値の対応関係

$$v(U) = \frac{P_L(U) \cdot W_L}{P_L(U) \cdot W_L + P_R(U) \cdot W_R}$$

但し，W_L と W_R は，条件を満たす記事数（N_L あるいは N_R）が多いほど大きくなるように設計された重みであり，以下の式で計算する．

$$W_L = \log_{10} N_L$$

$$W_R = \log_{10} N_R$$

以上の計算により得られる，単語 unigram U の印象語群 I_L に対する条件付き出現確率 $P_L(U)$ と印象語群 I_R に対する条件付き出現確率 $P_R(U)$ の重み付き内分比 $v(U)$ を，単語 unigram U の印象尺度「I_L ⇔ I_R」における印象値として印象辞書に登録する．

ここで，それぞれの印象尺度において印象辞書を構築するために用いた印象語群を表 2 に示す．これらの印象語群は，i）それぞれの印象尺度の印象を表す単語（動詞もしくは形容詞）であること，ii）語義の多様性により他の印象を（なるべく）持たない単語であること，という基準に基づいて決められている．

表 3 90 記事分のデータから生成された回帰式

印象尺度	回帰式（x：換算値）
楽しい ⇔ 悲しい	$-1.6355586x^3 + 18.971570x^2 - 70.68575x + 88.5147$
うれしい ⇔ 怒り	$2.384741939x^5 - 46.87159982x^4 + 363.6602058x^3 - 1391.589442x^2 + 2627.06261x - 1955.3058$
のどか ⇔ 緊迫	$-1.7138394x^3 + 21.942197x^2 - 90.79203x + 124.8218$

4.3 記事印象値の算出と 7 段階評価スケールへの換算

印象辞書を用いて新聞記事の印象値を算出する手法について述べる．

まず，4.1 節に示した方法で，入力された記事から単語 unigram を生成する．次に，生成された各単語 unigram の印象値を 4.2 節で構築した印象辞書から取り出し，印象尺度ごとに平均値を算出する．この平均値をその記事の当該印象尺度における印象値として扱う．なお，この印象値（以下，算出値と呼ぶ）は，印象尺度の左側の印象（楽しい，うれしい，のどか）が強いと 1 に近づき，右側の印象（悲しい，怒り，のどか）が強いと 0 に近づくように設計されているが，第 3 章で行ったアンケート調査では印象尺度の左側の印象が強いときは 1 に近づき，右側の印象が強いときは 7 に近づくという設計になっていたので，

$$\text{換算値} = (1 - \text{算出値}) \times 6 + 1$$

という式を用いて同じスケールになるよう算出値を換算した．

ここで，第 3 章で用意した全 90 記事（の第 1 段落）から求められる換算値とこの 90 記事（の第 1 段落）に対し回答者らが付けた 7 段階評価値の平均値（印象評価データ）の，それぞれの印象尺度における対応関係を散布図という形で図 1 に示す．

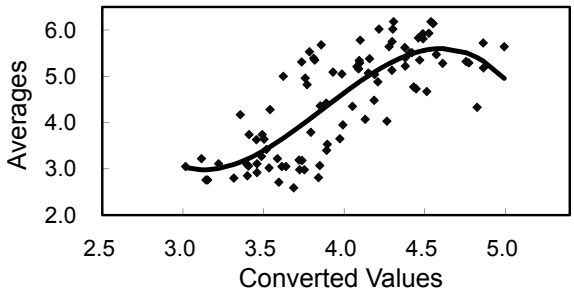
この図 1 に示した各散布図を見てみると，換算値と平均値の間には正の相関があることがわかる．そこで，各散布図における相関係数を計算してみたところ，上から 0.76，0.84，0.78 という高い数値であった．これは，印象辞書構築時に置いた「ある印象を有する単語 unigram は，その印象を表現する印象語群と共起しやすく，逆の印象を表現する印象語群とは共起しにくい」という仮定が全体的な傾向としては成り立っていることや換算値と平均値の対応関係を回帰分析により定式化できることを示唆している．

4.4 換算された印象値の回帰式による補正

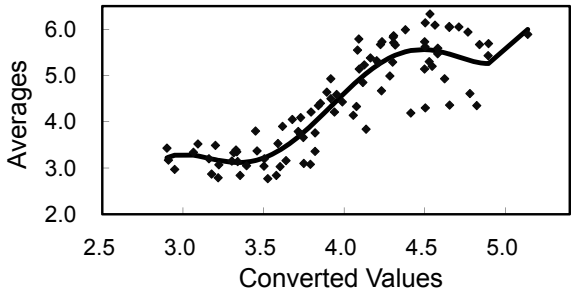
それぞれの印象尺度において，前節で得た換算値を説明変数，平均値を目的変数とする回帰分析を行い，両者の対応関係を示す最適な回帰式を得た．結果を表 3 に示す．この回帰式に換算値を代入することにより，換算値を補正する

表 4　回帰式の分析精度

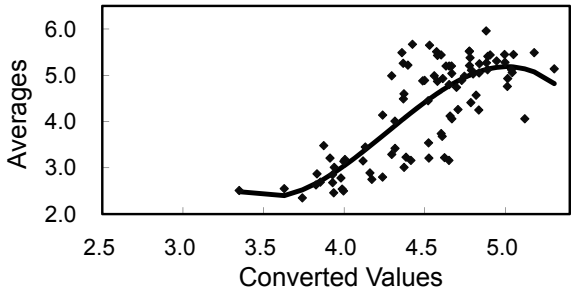
印象尺度	自由度修正済み決定係数
楽しい ⇔ 悲しい	0.62
うれしい ⇔ 怒り	0.79
のどか ⇔ 緊迫	0.63



(a)「楽しい ⇔ 悲しい」の場合



(b)「うれしい ⇔ 怒り」の場合



(c)「のどか ⇔ 緊迫」の場合

図 2　回帰分析の結果

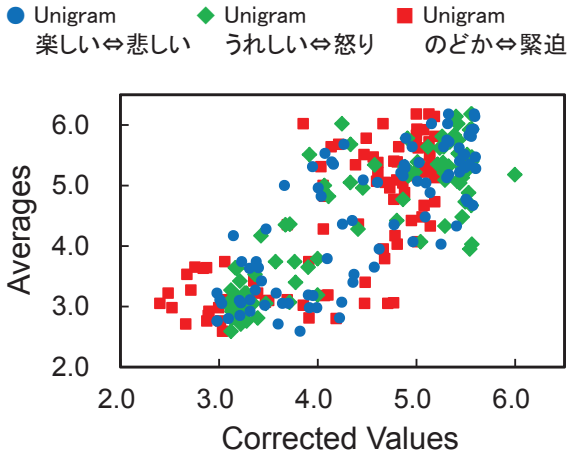
ことができる（以下，補正された換算値を補正値と呼ぶ）．

なお，この回帰分析では，様々な回帰モデル（直線，ロジスティック曲線，2 次関数，3 次関数，4 次関数，5 次関数など）が試され，その中から最も高い自由度修正済み決定係数 [13] を得たものが最適な関数として選ばれている．

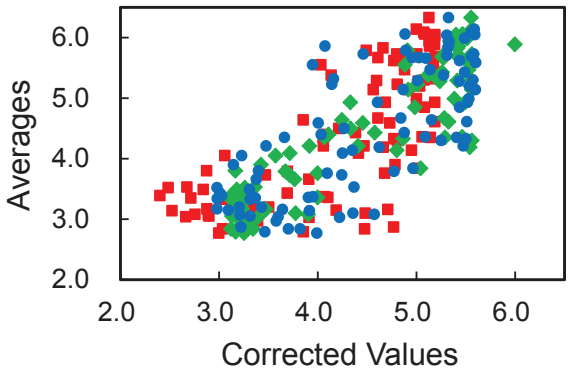
ここで，各回帰式の分析精度を表 4 に示し，各印象尺度における回帰分析の結果を図 2 にまとめる．表 4 によれば，いずれの印象尺度においても決定係数の値が 0.5 より高く，回帰分析の結果が良好であったことを示している．

5.　重回帰式による記事印象値の再計算

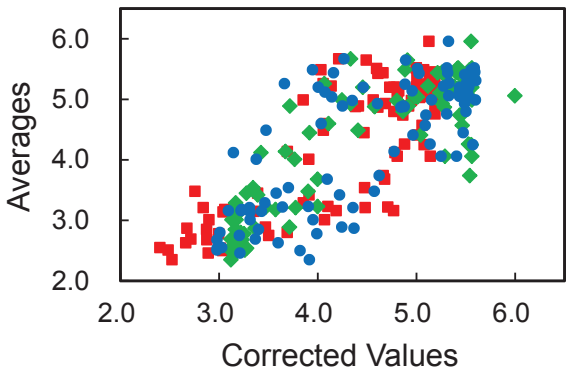
本章では，それぞれの印象尺度において求められた補正値を説明変数（3 個），アンケート調査に基づいて数値化された記事本来の印象値（回答者らが付けた 7 段階評価値の平均値）を目的変数とする重回帰分析を印象尺度ごとに行い，それぞれの対応関係を重回帰式という形で定式化する．このとき，変数選択法として変数増加法 [13] を採用することで，変数間の独立性が乏しいときに発生する多重共線性の問題を回避し，記事の印象値を求めるのに適した説明変数を取捨選択する．



(a)「楽しい ⇔ 悲しい」の場合



(b)「うれしい ⇔ 怒り」の場合



(c)「のどか ⇔ 緊迫」の場合

図 3　重回帰分析に資するデータの散布図

5.1　重回帰分析に資するデータの準備

まず，第 3 章で行ったアンケート調査の結果得られた各記事の各印象尺度における平均値を記事本来の印象値とみなし，重回帰分析における目的変数とする．一方，4.3 節で示した記事印象値の算出法と 4.4 節で示した記事印象値の補正法を用いて，回答者らに提示した記事（第 1 段落の

表 5 90 記事分のデータから生成された重回帰式

印象尺度	説明変数	偏回帰係数
楽しい	Unigram 楽しい⇔悲しい	0.313
⇔ 悲しい	Unigram うれしい⇔怒り	0.723
	（定数項）	（−0.152）
うれしい	Unigram うれしい⇔怒り	1.000
⇔ 怒り	（定数項）	（0.000）
のどか	Unigram うれしい⇔怒り	0.655
⇔ 緊迫	Unigram のどか⇔緊迫	0.383
	（定数項）	（−0.269）

表 6 重回帰式の分析精度

印象尺度	自由度修正済み決定係数
楽しい ⇔ 悲しい	0.73
うれしい ⇔ 怒り	0.80
のどか ⇔ 緊迫	0.75

み）から求められる 3 種類の補正値（3 種類の印象尺度に対応）を重回帰分析における説明変数とする．

ここで，全 90 記事から求めた 3 種類の補正値（3 個の説明変数）と回答者らが付けた 7 段階評価値の平均値（目的変数）との対応関係を印象尺度ごとに整理し，図 3 に示す．図中の点は，印象尺度（つまり，印象値の算出の際に用いた印象辞書の種類）によって区分されており，●印は「楽しい ⇔ 悲しい」に，◆印は「うれしい ⇔ 怒り」に，■印は「のどか ⇔ 緊迫」に対応している．なお，プロットが重なる場合は，●印が上に，■印が下になるように描かれている．

図 3 に示した散布図を見てみると，それぞれの補正値と平均値の間には正の相関があることがわかる．実際にそれぞれの相関係数を計算してみると，「楽しい ⇔ 悲しい」で 0.73〜0.85，「うれしい ⇔ 怒り」で 0.73〜0.90，「のどか ⇔ 緊迫」で 0.75〜0.85 という高い値であった．また，印象尺度によって分布の仕方に違いがあることもわかる．これらのことは，3 個の補正値と平均値との対応関係を重回帰分析により定式化することで，より高精度な手法を実現できることを示唆している．

5.2 重回帰分析に基づく重回帰式の生成

前節で準備したデータを用いて，3 種類の補正値を説明変数，回答者らが付けた 7 段階評価値の平均値を目的変数とする重回帰分析を印象尺度ごとに行った．その結果，表 5 に示す重回帰式が得られた．結局，ある新聞記事の，ある印象尺度における印象値は，対応する重回帰式に，その記事から算出される各説明変数の値を代入することにより，再計算される．例えば，ある記事の印象尺度「のどか ⇔ 緊迫」における印象値 y は，その記事から算出される説明変数「Unigram うれしい⇔怒り」，「Unigram のどか⇔緊迫」の値（補正値）をそれぞれ x_1，x_2 とすると，

$$y = 0.655x_1 + 0.383x_2 - 0.269$$

表 7 回帰式と重回帰式の導入による誤差（RMSE）の減少

印象尺度	換算値	補正値	出力値
楽しい ⇔ 悲しい	0.94	0.67	0.57
うれしい ⇔ 怒り	0.83	0.47	0.47
のどか ⇔ 緊迫	0.82	0.63	0.52

という式で算出される．なお，それぞれの重回帰式の自由度修正済み決定係数 [13] は，表 6 に示したように，いずれの印象尺度においても 0.5 より高く，重回帰分析の結果が良好であったことを示している．

6. 提案手法の有効性の検証

本章では，提案手法の学習データと未知データに対する精度を評価し，その有効性を検証するとともに，学習データに対する誤差解析を行い，今後の課題について考察する．

6.1 学習データに対する精度評価

まず，回帰分析と重回帰分析を行った際に用いた全 90 記事（の第 1 段落）を対象に，提案手法が出力する印象値と回答者らが付けた 7 段階評価値の平均値との誤差が回帰式や重回帰式の導入によりどう変化するかを調べた．但し，提案手法が出力する印象値は，3 種類あり，換算値（回帰式導入前の印象値），補正値（回帰式のみ使用時の印象値），出力値（回帰式・重回帰式併用時の印象値）となっている．結果を表 7 にまとめる．なお，誤差には，Root-Mean-Square Error（RMSE）を用いており，全 90 記事に対する平均値と換算値（もしくは補正値または出力値）の差分平方和を記事数で割り，平方根をとることにより求められる．また，このとき用いられた回帰式は表 3 に，重回帰式は表 5 に示されている．

表 7 によれば，すべての印象尺度において誤差が減少しているのがわかる．まず，回帰式を用いて換算値を補正することにより，誤差は，「楽しい ⇔ 悲しい」で 29.0 %，「うれしい ⇔ 怒り」で 42.7 %，「のどか ⇔ 緊迫」で 23.2 %減少しているが，重回帰式と併用することで，さらに減少し，最終的には「楽しい ⇔ 悲しい」で 39.7 %，「のどか ⇔ 緊迫」で 36.4 %の減少となっているのがわかる．「うれしい ⇔ 怒り」に対しては，重回帰式を導入した効果が見られないが，これは，表 5 に示した「うれしい ⇔ 怒り」の重回帰式が出力値＝補正値となっていることから当然と言える．しかしながら，その誤差は 0.47 と小さく，回帰式だけでも他の印象尺度より高い精度が実現されているのがわかる．また，「楽しい ⇔ 悲しい」や「のどか ⇔ 緊迫」の重回帰式を見てみると，その説明変数として「Unigram うれしい ⇔ 怒り」が取り込まれており，それぞれの印象尺度において高精度化に貢献しているのがわかる．これらのことは，回帰式を用いたときに高い精度を実現した説明変数「Unigram うれしい ⇔ 怒り」を，他の印象尺度でも利

表 8　5 分割交差検定による精度評価

印象尺度	換算値	補正値	出力値	従来手法 [8]
楽しい ⇔ 悲しい	0.94	0.70	0.60	0.69
うれしい ⇔ 怒り	0.83	0.49	0.49	0.49
のどか ⇔ 緊迫	0.82	0.62	0.52	0.64

用できるようにすることにより，高精度化が実現されていると言え，複数の説明変数を結合できる重回帰式を用いることの利点と言える．ここで，図 3 の散布図を見直してみると，それぞれの印象尺度において，補正値の分布範囲と平均値の分布範囲との間に若干のずれがあるのがわかる．回帰分析を行ったときの自由度修正済み決定係数 [13] は，「楽しい ⇔ 悲しい」で 0.62，「うれしい ⇔ 怒り」で 0.79，「のどか ⇔ 緊迫」で 0.63 となっており，ずれが最も小さかったのが「うれしい ⇔ 怒り」であることがわかる．このずれの小ささを重回帰式により取り込むことで，他の印象尺度でも誤差が減少したものと考えられる．

今回提案した第 2 のアプローチの長所は，説明変数の数を増やすことができる点にあり，印象尺度（すなわち印象辞書）を追加したり，高次の単語 N-gram を導入したり，新たな記事特徴量を導入したりすることで，更なる精度の向上も期待できる．

6.2　未知データに対する精度評価

次に，学習データを用いて 5 分割交差検定を行い，未知データに対する精度評価を行った．具体的には，(1) 90 記事分の学習データを 5 分割し，18 記事に対する換算値と平均値のデータセットを 5 つ作成する，(2) この 5 つのデータセットのうちの 4 つ（72 記事分の換算値と平均値）を用いて回帰分析と重回帰分析を行い，それぞれの印象尺度に対して最適な回帰式と重回帰式を構築する，(3) 残りのデータセット（18 記事分の換算値と平均値）を未知データとし，その換算値を回帰式に代入する，(4) その結果得られる補正値を重回帰式に代入し，提案手法の出力値（印象値）を得る，(5) 以上の処理の結果得られる，18 記事分の換算値，補正値，出力値と平均値との誤差（RMSE）を求める，という手順をすべての組合せ（5 通り）に対して行った．その結果得られた誤差（5 回分）の平均誤差を表 8 に示す．

まず，表 8 に示された未知データに対する誤差と表 7 に示された学習データに対する誤差を比べてみると，いずれの印象尺度においてもほぼ同等であることがわかる．特に，著者らが先行研究で提案した手法 [8] の誤差（表 8）と見比べてみると，「うれしい ⇔ 怒り」に対しては同じ精度を保ちつつ，「楽しい ⇔ 悲しい」と「のどか ⇔ 緊迫」に対する精度が大きく改善しているのがわかる．

以上の比較結果から，印象辞書を用いて算出される記事の印象値を回帰式を用いて補正するという文献 [8] で提案したアプローチと，それぞれの印象尺度において補正された印象値（補正値）を用いて各印象尺度における記事の印象値を算出し直すという今回提案したアプローチを組み合わせることで，未知データに対しても有効な，より高精度な手法を実現できることが確認された．

6.3　学習データに対する誤差解析

ここで，学習データ（全 90 記事分）を対象に，誤差（回答者らによる 7 段階評価値の平均値と提案手法による出力値の差の絶対値）が大きかった記事の内容を調べてみた．

まず，学習データから平均値と出力値の差が +1 以上であった記事を印象尺度ごとに抽出した結果，全部で 9 記事を得た．この 9 記事の各々の平均値は，4.89〜6.02 の範囲に分布しており，総じて負の印象（悲しい，怒り，緊迫）が強めであることがわかる．つまり，提案手法は，この 9 記事の印象を実際よりも弱く評価していたことになる．各記事の内容を調べてみると，その主な原因として，以下の二点に気付く．一つは，負の印象の強い単語があっても，そうでない単語が多いと，記事の印象（出力値）が弱められてしまうという点であり，これは，記事から抽出される特徴量の印象値を単純に平均している点に問題があるといえる．もう一つは，個々の単語にはあまり負の印象の強いものはないが，記事全体としては強い印象を感じる場合があるという点であり，これは，提案手法が個々の単語の印象値のみを処理対象とし，話題などの大局的な情報を取り入れていない点に問題があるといえる．

次に，逆の場合，すなわち平均値と出力値の差が −1 以下であった記事を印象尺度ごとに抽出し，11 記事を得た．この 11 記事の各々の平均値は，3.74〜4.36 に分布しており，中間値である「(どちらの印象も) 感じない (4 点)」に近い値となっている．これは，提案手法がこの 11 記事の印象を負の印象（実際には 5.04〜5.56 の範囲に分布）と判断したことを意味しており，平均値と出力値の差が +1 以上の場合とは逆のパターンになっている．そこで，各記事の内容を調べてみると，その主な原因として，一つのことに気付く．すなわち，負の印象の単語が比較的多く用いられている割に，記事の内容がさほど深刻ではないということである．例えば，某国の難民への支援物質を輸送した海上自衛隊の掃海母艦が母港に帰港したという話や，ダイヤモンドの原石をお腹の中にのみ込んでいた男が不法所持と密輸の疑いで逮捕されたという話，名誉毀損や安眠妨害に対して損害賠償を求める訴訟があったという話などが相当している．

以上，本節で述べたような問題の解決に際し，いくつかの側面からのアプローチを考える．まず，新たな説明変数の導入を検討する．例えば，記事から抽出される特徴量の印象値がどのように分布しているかを表す指標として，現在用いている平均値に加え，最大値や最小値，あるいは第

1 四分位数や第 3 四分位数といった統計量を用いることで，記事内における特徴量の印象分布を重回帰式に取り込むことが可能となり，高精度化に貢献することが期待される．次に，重要文抽出技術との組み合わせを考える．悲惨な状況下で起きた明るい出来事を伝える記事やその逆など正の印象の単語と負の印象の単語が混在する記事も見受けられることから，記事の印象を決定づけるような文（あるいは事象）を抽出した後，その文（あるいは事象）を対象に印象マイニングを行うという方法も有効かもしれない．また，話題のタイプによって人々の感じる印象が強くなったり，逆に弱くなったりすることもあるので，記事の話題タイプを決定する技術と組み合わせた上で，印象値を算出するための重回帰式を話題タイプごとに設計するという方法も考えられる．以上のようなことを今後の課題として取り組んでいきたい．

7. まとめ

本稿では，新聞記事を読んだ人々が感じる印象の強さを数値的に求めるための印象マイニング手法を提案した．本手法が対象とする印象は，「楽しい ⇔ 悲しい」，「うれしい ⇔ 怒り」，「のどか ⇔ 緊迫」の 3 種類であり，それぞれの印象に対して「(左側の印象を）感じる（1 点），割と感じる（2 点），やや感じる（3 点），(どちらの印象も）感じない（4 点），(右側の印象を）やや感じる（5 点），割と感じる（6 点），感じる（7 点）」という 7 段階の評価スケール（印象尺度）を設定している．提案手法は，それぞれの印象尺度において，このスケールに準じた 1.0〜7.0 の実数値を出力する．

具体的には，まず，著者らの先行研究 [8] をベースに，記事から抽出する特徴量として単語 unigram を定義し，新聞記事データベースから 3 つの印象辞書（3 種類の印象尺度に対応）を構築した．そして，それぞれの印象尺度において，印象辞書を用いて算出される記事の印象値を説明変数，アンケート調査に基づいて数値化される記事本来の印象値を目的変数とする回帰分析を行い，それぞれの対応関係を示す回帰式を構築した．この回帰式を用いることで，印象辞書を用いて算出される印象値が補正され，高精度な結果を得ることができるようになった．次に，それぞれの印象尺度において補正された 3 つの印象値を説明変数（3 個），記事本来の印象値を目的変数とする重回帰分析を行い，重回帰式を構築した．この重回帰式を用いることで，回帰式使用時に精度の良かった説明変数「Unigram うれしい ⇔ 怒り」を取り込むことが可能となり，その結果，さらに高精度な結果を得ることができるようになった．

提案手法の精度を，学習データ（全 90 記事分）を用いて評価してみたところ，重回帰式を導入して複数の印象辞書の相互利用を可能にすることにより，精度が大きく改善することが確認された．また，5 分割交差検定により未知データに対する精度を調べてみたところ，それぞれの印象尺度における平均誤差は，「楽しい ⇔ 悲しい」で 0.60，「うれしい ⇔ 怒り」で 0.49，「のどか ⇔ 緊迫」で 0.52 という結果であり，未知データに対する精度が学習データに対する精度とほぼ同じであること，従来手法 [8] より高精度であることが確認された．

今後の課題としては，1 つの記事の中での印象の推移を追跡できる印象マイニング手法の開発や，精度の向上に貢献する印象尺度（印象辞書）の設計・構築，印象の感じ方の違いを吸収する個人適応手法の開発，新たな説明変数の導入や他の技術（重要文抽出技術，話題タイプ決定技術）との統合などが挙げられる．一方，抽出された印象値を有効利用するアプリケーションの開発も必要不可欠であり，情報可視化 [14], [15] や情報推薦 [16], [17]，異メディアコンテンツ生成 [18], [19]，情報の信頼性評価 [20] といった様々な分野に応用していきたいと考えている．

謝辞 本研究の一部は，SCOPE 若手 ICT 研究者育成型研究開発（課題番号：102107001），福田将治奨学寄付金（52303）による助成，ならびに千葉工業大学附属総合研究所科学研究助成（平成 19 年度〜平成 24 年度）の成果であり，ここに記して謝意を表すものとする．

参考文献

[1] Peter D. Turney: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *ACL '02 Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417–424, Philadelphia, USA, 2002.

[2] Bo Pang and Lillian Lee: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. of the Annual Meeting on Association for Computational Linguistics*, pp. 115–124, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[3] Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen: Emotion classification of online news articles from the reader's perspective. In *Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 220–226, 2008.

[4] 宮川祥子，清木康：特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式，情報処理学会論文誌：データベース, Vol. 40, No. SIG 5 (TOD 2), pp. 15–28, 1999.

[5] 清水浩平，萩原将文：名詞と動詞の組み合わせに対する印象推定法，日本感性工学会論文誌, Vol. 10, No. 4, pp. 505–514, 2011.

[6] 阿部健一，藤本悠，大原剛三：ユーザーが受信メールから抱く感情に合わせた音楽推薦システム，第 4 回データ工学と情報マネジメントに関するフォーラム, No. A9-1, 2012.

[7] 秋山広美，小松孝徳，清河幸子：オノマトペから感じる印象の客観的数値化方法の提案，情処研報（ヒューマンコンピュータインタラクション研究会），第 2011-HCI-142 巻, pp. 1–7, 2011.

[8] 熊本忠彦，河合由起子，田中克己：新聞記事を対象とするテキスト印象マイニング手法の設計と評価，信学論（D），

Vol. J94-D, No. 3, pp. 540–548, 2011.
[9] Kenneth W. Church and Patrick Hanks: Word association norms, mutual information, and lexicography. *Computational Linguistics*, Vol. 16, Issue 1, pp. 22–29, 1990.
[10] 清木康，金子昌史，北川高嗣：意味の数学モデルによる画像データベース探索方式とその学習機構，信学論（D-II），Vol. J79-D-II, No. 4, pp. 509–519, 1996.
[11] CD-毎日新聞データ集 2002 年版.
[12] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao: Improvements of Japanese morphological analyzer JUMAN. In *Proc. of the International Workshop on Sharable Natural Language Resources*, pp. 22–28, Nara, Japan, 1994.
[13] 菅民郎：多変量統計分析，現代数学社，京都，2000.
[14] Tadahiko Kumamoto and Katsumi Tanaka: *Web OpinionPoll: Extensive Collection and Impression-based Visualization of People's Opinions*, Vol. 4, chapter 17, pp. 229–243. Springer, USA, 2008.
[15] 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.
[16] 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, Vol. 18, No. 2, pp. 173–183, 2006.
[17] Yukiko Kawai, Tadahiko Kumamoto, and Katsumi Tanaka: Fair news reader: Recommending news articles with different sentiments based on user preference. In *Proc. of International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Vol. LNAI 4692, pp. 612–622, 2007.
[18] 熊本忠彦，灘本明代，田中克己：記事の印象を伝達するニュース番組生成システム wEE の設計と評価，信学論（D），Vol. J90-D, No. 2, pp. 185–195, 2007.
[19] 石塚賢吉，鬼沢武久，加藤茂：物語のシーンの印象に基づいた声楽曲の生成，日本感性工学会論文誌, Vol. 10, No. 4, pp. 523–534, 2011.
[20] 山本祐輔，手塚太郎，アダムヤトフト，田中克己：WebAlert：Web 情報の印象集約を利用した閲覧ページ内容に対する反対意見提示，日本データベース学会論文誌, Vol. 7, No. 1, pp. 251–256, 2008.