1B1-1

回帰分析を応用したテキスト印象マイニング手法の設計と評価

Design and Evaluation of a Text-Impression Mining Method Using Regression Equations

熊本 忠彦[*1] 河合 由起子[*2] 田中 克己[*3]
Tadahiko Kumamoto Yukiko Kawai Katsumi Tanaka

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

This article focuses on the impressions which people feel from reading articles in newspapers and proposes a method for mining impressions of news articles. The target impressions to be mined are limited in this article to those represented by three impression scales consisting of two contrasting impression words, "Happy – Sad," "Glad – Angry," and "Peaceful – Strained," and the degree or strength of each impression is computed based on a seven-point scale from 1 to 7. First, our proposed method decomposes a news article into words, and obtains impression values for each of the words by consulting an impression dictionary. Next, the proposed method, for each impression scale, calculates the mean of all the impression values, and then corrects the mean value with the regression equation we have designed. Then the method outputs its corrected mean value as an impression value for the news article. Since it was considered that there were some gaps or errors between mean values calculated from news articles and the impressions which people felt from reading the news articles, we tried to narrow the gaps using regression analysis. We also verify the effectiveness of the proposed method based on a five-fold cross-validation, and prove that the accuracy of impression mining is highly improved by using regression equations.

1. まえがき

近年，インタラクションにおける感情の働きをモデル化し，コンピュータに感情を認識／表現させようという Affective Computing の概念 [1] が提唱され，感情の認識・伝達・利用に関する研究も盛んに行われている．しかしながら，誰かが発信したコンテンツを見たり聞いたりすることによって人が感じるであろう印象を抽出し，利用しようという研究は，まだ少ない．大辞林[*1]によれば，感情は「喜んだり悲しんだりする，心の動き．気持ち．気分」と定義されており，印象は「見たり聞いたりしたときに対象物が人間の心に与える感じ」と定義されている．すなわち，感情が情報発信者や情報受信者の心的状態もしくはその変化を意味し，深層的・心理的なものであるのに対し，印象は知覚レベルで作用する心象（イメージ）を意味しており，表層的・知覚的なものといえる．

そこで本稿では，新聞記事を対象に，情報発信者が発信したテキストから情報受信者が感じるであろう印象を抽出する，より高精度な印象マイニング手法を提案する．具体的には，「ある印象を有する単語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」という仮定のもと，新聞記事データを用いて，任意の単語と対比的な印象を有する 2 つの印象語群との共起の仕方を調べ，各単語が記事の印象に及ぼす影響を数値化した印象辞書を構築するとともに，この印象辞書を用いて新聞記事の印象値を算出するヒューリスティックな手法を開発する．さらに，このヒューリスティックな手法が新聞記事から算出する印象値と人々がその新聞記事を読んだときに感じる印象値の平均値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した印象値を補正するという方法で，より高精度な印象マイニングを実現する．また，提案手法の有効性を検証するために，被験者 100 人が新聞記事 10 記事を読み，それぞれの記事から受ける印象を評価するという実験を，被験者と記事を替えて 9 回行い，その結果得られる記事毎の印象値の平均値と提案手法が出力する各記事の印象値を比較することにより，精度評価を行う．なお，本稿において抽出対象となる印象は，「楽しい ⇔ 悲しい」，「うれしい ⇔ 怒り」，「のどか ⇔ 緊迫」の 3 種類であり，それぞれの印象軸に対し，「左側の印象を感じる（1 点）— わりと感じる（2 点）— やや感じる（3 点）— 感じない／わからない（4 点）— 右側の印象をやや感じる（5 点）— わりと感じる（6 点）— 感じる（7 点）」という 7 段階の評価尺度（印象尺度）が設定されている．すなわち，提案手法は，3 本の印象尺度のそれぞれにおいて，7 段階の評価値に準じた印象値（1.0〜7.0）を出力する．

連絡先: 熊本忠彦，千葉工業大学 情報科学部 情報ネットワーク学科，〒 275-0016 千葉県習志野市津田沼 2-17-1，kumamoto@net.it-chiba.ac.jp

*1 http://dictionary.goo.ne.jp/

2. 先行研究

我々はこれまで，テキストの印象というものに着目し，情報受信者がテキストからどのような印象を受けるかを推定する印象マイニング手法 [2] や，推定した印象を有効利用するためのシステム [3, 4] を開発してきた．Web 上のニュースサイトから記事を収集し，ニュース番組風の CG アニメーションを自動生成するシステム [3] では，記事の印象（明るい ⇔ 暗い）に応じて記事を読み上げる際の合成音声の声色（明るい声，普通の声，暗い声）や発話速度，声の大きさを変えることが可能になった．一方，文献 [4] の Web ニュース記事推薦システムでは，「明るい ⇔ 暗い」，「承認 ⇔ 拒否」，「緩和 ⇔ 緊張」，「怒り ⇔ 恐れ」という 4 つの印象尺度を構成することにより，管理可能なユーザの興味を話題だけでなく印象にも拡げ，複数のニュースサイトから収集した大量の記事を選択的に提示することが可能になった．しかしながら，これらのシステムのベースとなっている印象マイニング手法は，ヒューリスティックな手法であり，精度面での評価も十分にはなされていなかった．

3. 印象マイニング手法の設計と開発

3.1 設計方針

テキストを対象とする感情推定手法や印象マイニング手法では，テキストの構成要素である単語とその単語がテキストの感情／印象に与える影響との対応関係を表す辞書をいかにして構築するかがまず重要であり，先行研究ではそのような辞書を手作業で構築することもある（例えば文献 [5] など）．しかしながら，辞書を構築する際に作業者の判断を必要とする方法は，一般に高コストであり，i）テキストから受ける印象には個人差がある，ii）作業者の性格，体調，気分によって判断基準が変動する，iii）辞書の再構築や部分修正といったメンテナンスが容易でない，といった問題が生じることから，辞書の自動構築が必須となってくる．

そこで，我々が文献 [2] で提案したように，何らかのヒューリスティックな知識を導入し，機械的な作業のみで辞書を自動構築できるようにすることを考える．しかしながら，このようなヒューリスティックな方法だけでは，自動的に算出された印象値と人々が感じる印象値との間にギャップが生じてしまうと考えられることから，本稿では，算出された印象値を補正することにより，人々が感じる印象値とのギャップを狭めることを考える．すなわち，新聞記事から算出された印象値と人々がその新聞記事を読んだときに感じる印象値との対応関係を回帰分析により求め，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ことにする．

3.2 印象辞書の自動構築

本稿では,「ある印象を有する単語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[*2] のもと，新聞記事データを用いて，任意の単語と対比的な印象を有する 2 つの印象語群との共起の仕方を調べ，数値化したものを，その単語の印象値として印象辞書に登録する．すなわち，5 年分の読売新聞記事データ（2002 年版〜2006 年版）を解析することにより，3 本の印象尺度「楽しい ⇔ 悲しい」,「うれしい ⇔ 怒り」,「のどか ⇔ 緊迫」のそれぞれに対して 0〜１.0 の実数値が求められ，印象値として印象辞書に登録される．具体的な手順を以下に示す．

まずはじめに，各印象尺度の左側の印象を表す印象語群 IW_L と右側の印象を表す印象語群 IW_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 IW_L あるいは IW_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．但し，何らかの理由で印象語が列挙されていても極端に重要視されないよう，同じ印象語群の印象語が 3 語以上含まれている場合は 3 語として扱うことにする．以上の結果，印象語群 IW_L に属する印象語の数が印象語群 IW_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）から助詞，連体詞，指示詞以外のすべての単語を抽出し，単語ごとに出現頻度を数える．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，

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

と表される．但し，$P_L(w)$ ／ $P_R(w)$ は，印象語群 IW_L ／ IW_R に属する印象語がある記事で用いられたときに，単語 w もその記事で用いられる条件付確率を補正したものであり，印象語群 IW_L ／ IW_R に属する印象語の数が印象語群 IW_R ／ IW_L に属する印象語の数よりも多かった記事を対象とすることにより，仮定に即した記事のみを利用している点が文献 [2] で用いられた条件付確率とは異なる．

この $P_L(w)$ と $P_R(w)$ を用いて，単語 w の印象値 $v(w)$ を次のような式で表す．

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

$$weight_L = \log_{10} N_L, \qquad weight_R = \log_{10} N_R$$

すなわち，単語 w の印象語群 IW_L に対する補正済み条件付確率と印象語群 IW_R に対する補正済み条件付確率の重み付き内分比を求め，単語 w の印象尺度「IW_L ⇔ IW_R」における印象値として印象辞書に登録する．なお，$weight_L$ と $weight_R$ は重みであり，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ている．

印象辞書構築の際に，それぞれの印象尺度に対しシードとして与えた印象語群を表 1 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，という基準の下，若干の試行錯誤に基づいて決定された．なお，本手法では形態素解析システムとして JUMAN[6] を用いているが，JUMAN の出力をそのままの形では用いていない．複雑かつ細かい処理なので，詳細は割愛するが,「削除しない」のような語をサ変名詞「削除」，動詞「する」，形容詞性述語接尾辞「ない」の 3 語に分けるのではなく，動詞 1 語として扱うためのルールや,「ホームランだ」のような語を普通名詞「ホームラン」と判定詞「だ」に分けずに，形容詞 1 語として扱うためのルール,「再チャレンジ」のような語を名詞接頭辞「再」とサ変名詞「チャレンジ」に分けずに，サ変名詞 1 語として扱うためのルールなどが用意されており，印象辞書構築時ならびに記事印象値算出時に JUMAN の出力（形態素解析結果）に対して適用されている．

表 1: 各印象尺度を構成する印象語群

印象尺度	印象語
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

3.3 記事印象値の算出と回帰式による補正

新聞記事 A に含まれている単語の印象値を印象辞書から取得し，その平均値を計算する．この平均値が記事 A の印象値 $O(A)$ となるわけだが，3.1 で述べたように，記事から算出される印象値 $O(A)$ とその記事に対して人々が感じる印象値との間にどのような対応関係があるのかは考慮されていない．そ

*2 文献 [2] でも同様の仮定を置いてヒューリスティックな手法を開発したが，各印象尺度を構成するのは 2 つの印象語群ではなく，印象語 2 語だけであった．なお，この印象マイニング手法を他のアプリケーション [3, 4] に応用する際に，対比的な 2 つの印象語群から 1 本の印象尺度を構成できるように改良したが，改良後の精度評価は行っていなかった．

表 2: 各印象尺度の印象値を計算するための回帰式

印象尺度	回帰式
楽しい ⇔ 悲しい	$-1.636x^3 + 18.972x^2 - 70.686x + 88.515$
うれしい ⇔ 怒り	$2.385x^5 - 46.872x^4 + 363.660x^3 - 1391.589x^2 + 2627.063x - 1955.306$
のどか ⇔ 緊迫	$-1.714x^3 + 21.942x^2 - 90.792x + 124.822$

表 3: 回帰式の精度評価

印象尺度	決定係数	重相関係数
楽しい ⇔ 悲しい	0.63	0.79
うれしい ⇔ 怒り	0.81	0.90
のどか ⇔ 緊迫	0.64	0.80

こで本稿では，新聞記事 90 記事を対象に，提案手法が算出した印象値と被験者 100 人が感じた印象値の平均値を用いて回帰分析を行い，その対応関係を定式化する．

まずはじめに，新聞記事とその記事から受ける印象との対応関係を示す正解データを得るために，被験者 900 人（男女それぞれ 450 人ずつ）を対象に印象評価実験を行った．具体的には，被験者 900 人を 9 つのグループ（男女 50 人ずつ，計 100 人）に分け，各グループに毎日新聞の 2002 年版社会面[*3]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た．各被験者は，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 本の印象尺度（7 段階）を用いて評価した．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）— 感じない／わからない（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，正解データとした．なお，各被験者に提示された記事は，元の記事の第 1 段落のみであり，個人情報保護の観点から個人の特定につながりうる情報（個人名，組織名[*4]，地域名[*5]）を記号（◇や○，□）と置換し，伏せ字とした．

次に，3.2 で構築した印象辞書を用いて，被験者らに提示された各記事（第 1 段落）の各印象尺度における印象値を算出した．なお，算出値は，印象尺度の左側の印象が強いと 1 に近づき，右側の印象が強いと 0 に近づくように設計されているが，印象評価実験では印象尺度の左側の印象が強いときは 1，右側の印象が強いときは 7 という設計になっていたので，

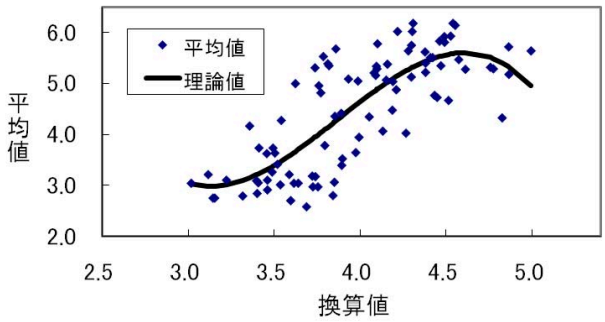
$$\text{換算値} = (1 - \text{算出値}) * 6 + 1$$

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．以下の回帰分析では，この換算値を用いる．

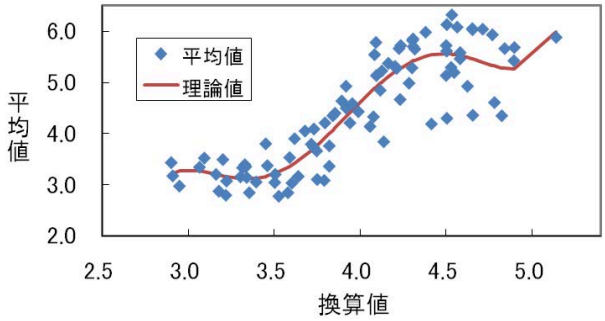
*3 印象辞書の構築には 2002 年版～2006 年版の 5 年分の読売新聞記事データを用いたが，評価実験では毎日新聞 2002 年版（社会面）の記事データのみを用いた．

*4 公的・公共機関の所属名は置換しないで，部署名のみを置換した．但し，所属名に地域名が含まれているときは，その部分も置換した．大学・企業・団体等の場合は,「大学」や「会社」など組織の種別を表す部分のみ置換しなかった．

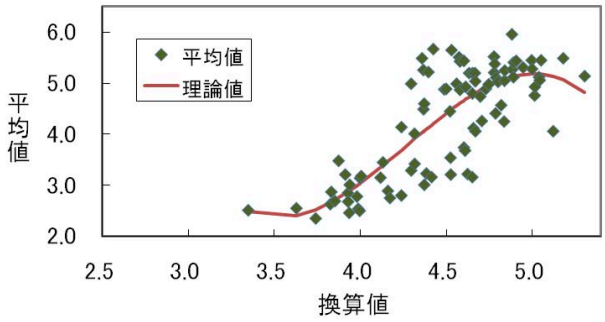
*5 「県」や「市」など地域種別を表す部分は置換しなかった．



(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: 回帰分析の結果

以上の結果，新聞記事 90 記事に対して，提案手法が算出した印象値の換算値と被験者 100 人が付与した印象値の平均値が得られた．そこで，これらのデータを学習データとし，様々な回帰モデルを用いて回帰分析 [7] を行ったところ，表 2 に示す回帰式の決定係数が最も高く，両者の対応関係を示す最適な関数として選ばれた．したがって，この回帰式に換算値を代入することにより，記事の印象値が出力されることになる．ここで，学習データに対する回帰式の精度を表 3 にまとめ，学習データの散布図と回帰分析の結果得られた回帰式を図 1 に示す．表 3 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ることを示している．

4. 評価実験

本章では，提案手法の精度評価を行い，その有効性を検証する．

まず，回帰式の構築時に用いた全 90 記事を対象に，提案手

表 4: 回帰式の導入による誤差（ユークリッド距離）の変化

印象尺度	導入前	導入後	改善率	ベース
楽しい ⇔ 悲しい	0.94	0.67	29.0%	0.99
うれしい ⇔ 怒り	0.83	0.47	42.7%	0.82
のどか ⇔ 緊迫	0.82	0.63	23.2%	1.00

表 5: 5 分割交差検定による精度評価

印象尺度	平均	最大値	最小値
楽しい ⇔ 悲しい	0.69	0.78	0.57
うれしい ⇔ 怒り	0.49	0.58	0.42
のどか ⇔ 緊迫	0.64	0.81	0.50

法が出力する印象値[*6] と被験者らが付与した印象値の平均値の間の誤差が回帰式を導入する前と後でどう変化するかを調べてみた．結果を表 4 にまとめる．なお，誤差は，全 90 記事に対する換算値／理論値と平均値の差分平方和を記事数（= 90）で割り，平方根をとったものであり，いわゆるユークリッド距離となっている．また，改善率は，

$$\text{改善率} = \frac{\text{導入前の誤差} - \text{導入後の誤差}}{\text{導入前の誤差}}$$

と定義されている．一方，ベースライン手法は，文献 [2] で提案した手法をベースにしており，印象辞書を構築する際には印象語群ではなく印象語 2 語[*7] が用いられているが，解析対象となる新聞記事データには提案手法と同じもの（読売新聞記事データ 2002 年版〜2006 年版）を用いている．

表 4 によれば，すべての印象尺度において比較的高い改善率を示しており，回帰式を導入することにより，精度が向上することが確認された．特に「うれしい ⇔ 怒り」に対しては，誤差が 0.5 未満となっており，高い精度を達成している．

以上の結果，学習データに対しては，十分な結果を得ることができた．そこで以下では，未知データに対する精度評価を 5 分割交差検定により行う．すなわち，(1) 正解データを 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）を用いて回帰分析を行い，それぞれの印象尺度に対して最適な回帰式を構築する，(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，換算値を回帰式に代入することにより，理論値を得る，(4) この理論値と平均値の誤差を計算する，という手順を任意のデータセットに対して行った．結果を表 5 にまとめる．表 4 の「導入後」の誤差と表 5 の「平均」の誤差を比べてみると，ほぼ同等であり，未知データに対しても十分有効であることがわかる．

5. むすび

本稿では，新聞記事を読んだ人々がその記事から感じるであろう平均的な印象を抽出する手法を提案した．具体的には，「ある印象を有する単語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[*8] を置き，読売新聞記事データ 2002 年版〜2006 年版の 5 年分を用いて，記事中に現れる各単語と対比的な印象を有する 2 つの印象語群との共起関係を数値化し，その単語の印象値として印象辞書に登録する手法を開発した．さらに，この印象辞書を用いて求められる新聞記事の印象値と人々がその記事を読んだときに感じる平均的な印象値との対応関係を回帰分析により調べ，回帰式という形で定式化することにより，記事の印象をより高精度に抽出する印象マイニング手法を提案した．

今後の課題としては，今回採用したアプローチがどのような印象尺度に対し有効なのか，印象語群をどう選べば印象マイニングの精度を向上させることができるのか，といったことを検討・考察していくことが挙げられる．また，音楽や絵画といったマルチメディアを対象とする印象マイニング手法では，個人差の問題が重要な研究トピック [8] の一つとなっており，テキストを対象とする印象マイニング手法でも同様に，対処していく必要がある．

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] R. W. Picard, Affective Computing, MIT Press, 1997.

[2] T. Kumamoto, and K. Tanaka, Proposal of Impression Mining from News Articles, Lecture Notes in Artificial Intelligence, LNAI3681, KES2005, Springer, pp. 901–910, 2005.

[3] 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.

[4] 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.

[5] 松本和幸，黒岩眞吾，任福継，感情計測システムについて，信学技報，NLC2003-10，pp.55–60，2003.

[6] 黒橋禎夫，河原大輔，日本語形態素解析システム JUMAN version 5.1，`http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html`，2005.

[7] 菅民郎，多変量統計分析，現代数学社，京都，2000.

[8] 熊本忠彦，印象に基づく楽曲検索のためのユーザモデリング手法，情処学論：データベース，Vol.47，No.SIG 8 (TOD 30)，pp.157–164，2006.

*6 回帰式を導入する前は換算値が印象値であり，導入後は回帰式に基づいて算出された理論値が印象値となる．

*7 「楽しい ⇔ 悲しい」には「楽しい」と「悲しい」を，「うれしい ⇔ 怒り」には「うれしい」と「怒る」を，「のどか ⇔ 緊迫」には「のどかだ」と「緊迫する」を用いた．

*8 新聞記事データベースを調べてみると，この仮定が成り立っていない記事を見つけるのはさほど難しくないことがすぐにわかるが，その一方で，全体的な傾向としては成り立っているようにも感じられる．実際，図 1 に示した散布図において，換算値と平均値の間には正の相関があることが見て取れ，仮定の妥当性を示している．なお，各散布図における相関係数は 0.76〜0.84 であった．