Webドキュメントを対象とした情緒的表現解析システムの試作

A prototype system for analyzing emotional expressions of Web documents

徳久 雅人[*1]
Masato Tokuhisa

[*1] 鳥取大学大学院工学研究科情報エレクトロニクス専攻
Department of Information and Electronics, Graduate School of Engineering, Tottori University

People's concerns are described in Web documents, especially "blogs." Since affect represents the concerns in the abstract, it is hopeful to analyze emotional expressions of blogs in order to extract useful information for marketing researches, public services, etc. In this paper, a prototype system for the emotional analysis is constructed and evaluated. The system contains "RSS receiver," "blog downloader," "text extractor," and "emotion reasoner." The RSS receiver watches the blog sites, and the blog downloader gets the articles. The text extractor parses HTML to distinct main text part and scans the text to split/concatenate the lines by a sentence unit. The emotion reasoner infers emotions and the targets of them based on valency patterns. The experiments show that the accuracy of the emotion reasoning is the same level as manually reasoning, and the targets of emotions give good hints for the concern analysis.

1 はじめに

近年，多くの人々が World Wide Web を通じて興味・要望・不満などの情報を直接的あるいは間接的に発信している．これらはマーケティングや行政的活動に役立つ情報であるので，その自動抽出技術が期待されている．興味や不満などの主観性を抽象的に表すものとして情緒・感情が挙げられるので，Webドキュメントから情緒的表現を解析することで，Web に展開されている主観的な情報を鳥瞰することがきると予想される．そこで，本稿では，主観的な情報を幅広く見渡すために，Webドキュメントから情緒的な表現を解析することを，目的とする．

先行研究として，既に Webドキュメントの収集やテキストからの情緒推定は行われている．Webドキュメントの収集について，関根らは Yahoo!カテゴリを基に収集を行い，南野らは様々な Blog サイトからの自動抽出を行った[関根 05][南野 04]．ここで，最近の Blog サイトは，新しい記事の投稿を RSS で発信するサービスが備わっている．時系列に沿って情報を収集する上では，RSS に基づく Blog サイトからの記事の取得が有効と思われる．一方，Webドキュメントの解析について，Kanayama らは要望を表すパターン知識を用いて製品のニーズを解析した[Kanayama 08]．情緒に関する解析では，目良らは情緒生起の原因を表す文を深層格フレームと好感度計算式を用いて解析した[目良 08]．これらより，パターン知識は情報の抽出に有効であること，および，情緒の解析は用言ごとの判定が必要であることがわかる．そこで，本稿では，RSS に基づく Blog 記事の取得を行い，パターン知識を用いた情緒の解析を行うという手法をとる．

本稿は，第 2 章で情緒推定の原理と解析システムの構成を述べる．第 3 章で，Webドキュメントの取得した結果を示す．第 4 章では，取得したドキュメントから情緒的表現を解析した結果を示す．第 5 章で考察を述べ，第 6 章でまとめる．

2 情緒的表現解析システム

2.1 原理

情緒には過程がある．すなわち，生じる原因があり，その結果，ある情緒状態になり，時として情緒的な表出や反応が生じる．そこで，本稿での情緒的表現とは，テキストの書き手が，情緒の生じる原因となる状況に置かれている表現，情緒状態や情緒的な表出を明示的に表す表現のことを指す．

以下にそれぞれの例文を示す:

例 1 大切なものを無くした。(情緒原因を表す文)

例 2 左折車に腹が立った。(情緒状態を表す文)

例 3 その知らせで涙を浮かべた。(情緒的表出を表す文)

これらの文から情緒的な情報を解析するには，次のパターンが有効である．

パターン 1. *N*1 が *N*2 を無くす
　過程名: 原因, 情緒名: 悲しみ, 情緒主: *N*1, 情緒対象: *N*2

パターン 2. *N*1 が *N*2 に腹を立てる
　過程名: 状態, 情緒名: 怒り, 情緒主: *N*1, 情緒対象: *N*2

パターン 3. *N*1 が *N*2 で涙を浮かべる
　過程名: 表出, 情緒名: 悲しみ, 情緒主: *N*1, 情緒対象: *N*2

パターンが適合することにより，それぞれの情緒名を挙げるとともに，情緒対象が，例 1 では「大切なもの」，例 2 では「左折車」，例 3 では「その知らせ」ということが抽出できる．

本稿では，「過程名」，「情緒名」，「情緒主」，「情緒対象」というスロットおよびその値を「情緒属性」と呼ぶ．過程名には「原因」，「状態」，「表出」および「非情緒」という値をとる．情緒名には，9 分類系の情緒名をとる．具体的には，《喜び》，《好ましい》，《期待》，《悲しみ》，《恐れ》，《嫌だ》，《怒り》，《驚き》，《なし》である．情緒主と情緒対象は，パターンの変数が対応する．

ここで，情緒名の細かさに関して，5 分類系と 3 分類系を用意している．これらは 9 分類系から単純に求まるものである．すなわち，5 分類系の情緒名は，《 P 》，《 N 》，《 A 》，《 S 》，《なし》である．《 P 》は《喜び》，《好ましい》，《期待》に，《 N 》は《悲しみ》，《恐れ》に，《 A 》は《怒り》，《嫌だ》に，そして，《 S 》は《驚き》にそれぞれ対応する．3 分類系の情緒名は，《 Positive 》，《 Negative 》，《なし》である．《 Positive 》は《 P 》に，《 Negative 》は《 N 》，《 A 》に，そして，3 分類系の《なし》は《 S 》と 5 分類系の《なし》に，それぞれ対応する．

2.2 構成

本システムは，「 RSS 受信部」，「ブログダウンロード部」，「テキスト抽出部」，および，「情緒推定部」で構成する．

連絡先:徳久雅人，鳥取大学大学院工学研究科情報エレクトロニクス専攻，〒 680-8552 鳥取市湖山町南 4-101，tokuhisa@ike.tottori-u.ac.jp

RSS 受信部では，特定のブログサイトからの RSS を受信する．RSS は 30 秒～ 10 分の周期で受信する．更新頻度に応じて周期を変更する．

ブログダウンロード部では，RSS 取得後 24 時間以上あけて，ブログ記事をダウンロードする．コメントを取得すること，および，一時的な掲載記事の取得を避けること，という目的がある．

テキスト抽出部は，HTML ソースファイルから，ブログ本文を抽出すること，および，テキストから 1 文ずつを抽出することを行う．ブログ本文の抽出は，HTML ソースファイルの解析によるものであり，手作業で作成したルールを用いる．これはブログサイトごとに用意する．文の抽出では，複数の文から，文の境界を検出する．ブログでは，句読点が省略されることがあるためである．特定の文字に基づく規則の他，句読点の前に出現する文字列の出現確率のテーブルを使用する．テーブルには，181,807 種類の文字列が収録されている．これらは，ブログ記事 4 日ぶんから作成した．

情緒推定部では，パターン辞書を使用する．パターン辞書は，日本語語彙大系の文型パターンがベースとなっており，情緒原因，情緒状態，情緒表出についての属性が付与されている[池原 97][田中 04][黒住 06]．また，情緒と関係しないことも記されている．文型パターン数は，14,800 件であるが，1 パターンに複数の属性セットが付与されている（1 つの属性セットは，第 2.1 節で示したように，過程名，情緒名，情緒主，情緒対象で構成）．情緒の原因については 11,724 セット，情緒の状態については 1,035 セット，情緒の表出については 129 セット，非情緒については 6,923 セットが含まれている．

3 Web ドキュメントの取得

3.1 RSS 受信とブログダウンロード

ブログサイト 3 社から RSS を受信した．2008 年 8 月 1 日から 2009 年 1 月 23 日までの 147 日間取得したところ，6,816,633 件の記事を得た．ブログサイトによるのだが，1 日平均で 9 千～2.4 万件が RSS で発信されていたことが分かった．なお，この期間に建物工事などの外的要因のため取得していない日があったが，その期間を除き，連続して稼働できることが確認できた．

3.2 テキスト抽出

まず，ブログの主要部分を HTML ソースファイルから抽出することは，手作業で作成したルールでほぼ安定して処理することができた．ルールは正規表現を主とするもので十分であった．

次に，記事から文を抽出する．以下に例を示す．

例 4

（原文）

1: 今日は陽気もよく、始めてすぐ汗が出てきたし

2: ひねり運動で腰がイイ感じに疲れましたね～

3: ウエスト引き締め効果もあるかも…？

（文抽出結果）

1: 今日は陽気もよく、始めてすぐ汗が出てきたしひねり運動で腰がイイ感じに疲れましたね～

2: ウエスト引き締め効果もあるかも…？

原文 1 行目の末尾が文末でないと判定されて，原文 2 行目と結合されていることが，文抽出結果より確認できる．

ランダムで抽出した記事集を 2 セット準備して，文抽出の精度を確認した．テストセットは，1 つ目が 2,299 文，2 つ目が 1,420 文を入力文として用意した．提案手法は，ヒューリスティクスと句読点前確率を用いた判定である．

比較のために 2 つのベースライン手法を用いた．ベースライン手法 1 は，文分割を行わない手法である．ブログ記事の 1 行を 1 文とみなす方法である．ベースライン手法 2 は，「。」，「、」，「！」，「？」に基づいて判定する方法である．

結果は表 1 のとおりである．提案手法が良好に動作したことが確認できた．なお，誤りの理由は，連体修飾節を文末であると誤判定することである．

表 1 ブログ記事からの文抽出の精度

手法	テストセット 1	テストセット 2
提案手法	89.7%	92.5%
ベースライン手法 1	70.8%	67.8%
ベースライン手法 2	66.9%	71.1%

4 情緒的表現の解析

4.1 解析の様子

ブログ記事を入力すると，本システムは，文脈参照せずに文単位で情緒推定を行い，情緒属性を出力する．以下に具体例を示す．

例 5

INPUT=AA000001

本日+8+時+に/出発し+て/芦別+で/焼き+たて/りんご+パイ+を/買う+。

[5, 出発し, 原因, emotion:期待, feeler: φ , feelto: φ]

[5, 出発し, 原因, emotion:恐れ, feeler: φ , feelto: φ]

[9, 焼き, 原因, emotion:期待, feeler: φ , feelto: φ]

[9, 焼き, 原因, emotion:喜び, feeler: φ , feelto: φ]

[14, 買う, 原因, emotion:喜び, feeler: φ , feelto:りんごパイ]

例 6

INPUT=AA001270

自分+を/信じ+てい+た+から

[3, 信じ, 状態, emotion:期待, feeler: φ , feelto:自分]

例 7

INPUT=AA001118

電話+って+何+だ+か/緊張する/んだ+よ/なぁ+・+・

[6, 緊張する, 表出, emotion:期待;恐れ, feeler: φ , feelto: φ]

例 5 では，「出発し」より《期待》と《恐れ》が，「焼き」より《期待》と《喜び》が，そして，「買う」より《喜び》がそれぞれ推定されている．情緒対象は「りんごパイ」である．いずれも，情緒過程名が「原因」である．これより，ブログの著者が自身の情緒を明示しているわけではないが，情緒の生じる可能性があるものとして解析されている．例 6 は，情緒過程名が「状態」である．「自分」に対する《期待》があったことが明示されている．例 7 は，情緒過程名が「表出」である．「緊張する」という身体変化が生じていることから，《期待》や《恐れ》がその原因であるとして，情緒が推定されている．なお，情緒対象が明示されていない場合，および，抽出できなかった場合には，「φ」が出力される．

4.2 性能の評価

本システムは，情緒の推定されたブログの文から情報を抽出する．したがって，本システムが，「情緒有り」と判定した範囲での正しさが評価されるべきである．ここで，知識ベースを構築する際，9 分類系を採用したが，情報を抽出する上で，たとえば，《喜び》と《好ましい》の区別があまり重要でない．ゆえに，情緒の種類は，5 分類系や 3 分類系を採用し，正解とする情緒との一致を評価する．正解について，人間 5 名（男子大学 4 年生）

により情緒推定を行い，1 名でも推定した情緒であれば正解とする．出力と正解の一致は次式で計算する．

$$〈一致率〉 = \frac{2N(o \cap c)}{N(o) + N(c)}$$

$N(x)$ は集合 x の要素数，o は出力された情緒の集合，c は正解の情緒の集合，$o \cap c$ は同一文において出力と正解の一致した情緒の集合を，それぞれ表す．

評価実験に用いた文は，307 文である．9 分類系，5 分類系，3 分類系での各一致率を表 2 に示す．

表 2 情緒推定の一致率

分類系	一致率
9	0.375
5	0.592
3	0.685

予備実験において，人間の間での同様の一致率を調べたところ，9, 5, 3 分類系のそれぞれで，0.513, 0.566, 0.618 であった．したがって，5 および 3 分類系の自動推定に対しては概ね良好であると言える．

4.3 情緒対象の分布

(1) 目的と方法

ブログ記事を対象に，情緒対象を基準とした情報の分布を調査する，この調査により，人々の関心や不満の対象が観測できることを期待する．

調査の方法は，まず，記事を文単位で情緒推定し，情緒の過程名，情緒名，情緒対象を抽出する．次に，情緒名ごとに，情緒対象の出現頻度を求める．普遍的に出現する情緒対象を取り除き，情緒対象として注目されているキーワードを調べる．

(2) 結果

調査対象は，あるブログサイトの 2008 年 8 月に RSS 発信された記事 4 日ぶんである．記事本文は 124 万文であった．

まず，過程名に対応する表現の件数を調べると，表 3 のとおりとなった．非情緒過程を表すと解析された箇所は，1 文あたり平均 0.42 箇所であり，情緒過程を表すと解析された箇所は，1 文あたり平均 0.68 箇所であった．情緒過程の中でも，情緒の原因を表す箇所は，情緒の状態・表出を表す箇所よりも多かった．一般に，情緒の状態と表出に着目する研究事例が見られるが，原因に着目することでより広範囲から情報の抽出が可能であることがわかる．

表 3 情緒過程の表された箇所

過程名	箇所数
非情緒過程	522,390
情緒過程	
原因	688,819
状態	143,808
表出	5,828
合計	1,360,845

次に，情緒過程とされた表現における情緒名の頻度を，表 4 にまとめる．前節より情緒推定の精度は，5 分類系の粒度において信頼できるものであるので，ここでは，5 分類系に基づく件数をまとめる．この表より，《P》(喜び，好ましい，期待)の情緒は，《N》(悲しみ，恐れ)と《A》(嫌だ，怒り)よりも多いが，《N》と《A》の合計と同程度の頻度であることが分かる．

表 4 情緒名の対応付けられた件数

情緒名(5 分類系)	件数
《P》	632,199
《N》	302,447
《A》	257,296
《S》	9,819

抽出された情緒対象(φを含む)は，73,706 種類であった．出現頻度の順でみると，上位 5 件に「φ(899,775 回)」，「元気】(486 回)」，「【大卒フリーター】(427 回)」，「オンライントレード】(346 回)」，「少しホットミルク(334 回)」が得られた．しかし，「φ」は，いずれの情緒にも出現している．総出現頻度の順位付けでは，特定の情緒に現れる情緒対象が高い順序に見られなくなる．そこで，特定の情緒について対象になる確率が高く，全体の情緒についての対象になる確率が低いものを表すスコアを付けることにする．次式で計算する．

$$s(w) = P_e(w) \log^3 P_a^{-1}(w)$$

ここで，w は情緒対象，$s(w)$ は w のスコア，$P_e(w)$ は指定する情緒名での w の出現確率，$P_a(w)$ は w の総出現確率である．なお対数の冪数は，予備実験により定めた．

表 5 に，情緒ごとに抽出された情緒対象をまとめる．情緒対象をキーワードとしてブログ記事を検索すると，関心の高い事柄であることがわかる．たとえば，「豆乳」はダイエットや調理用の食材として，「ファンタシースターポータブル」は人気ゲームソフトとして，注目されていることがわかる．しかし，SPAM と思える記事から抽出されているものもある．たとえば「広告費」については，繰り返し同じ文面が出現した．

5 考察

情緒対象をキーワードとしてブログ記事を検索してみると，記事全体から読み取られる情緒は，表 5 で参照した際の情緒とは異なることがある．情緒対象に関連する情緒名をより正確に解析することが今後の課題となる．

たとえば，情緒《S》(驚き)の対象として高い順位に出現した「中毒事件」は，情緒対象として 11 箇所から抽出されている．しかし，ブログ記事を参照すると，120 文が存在し，記事全体からは，この対象に対して，《怒り》や《恐れ》などが読み取られる．ゆえに，情緒対象を抽出した後に，その情緒対象を含む文から推定した情緒を，情緒対象に関連する情緒情報として再利用する方法が考えられる．

6 おわりに

Web ドキュメントを RSS 受信に基づき自動収集し，情緒的表現を解析するシステムを試作した．Web ドキュメントの取得は，半年間の連続運転において動作が確認できた．情緒的表現の解析は，5 分類系の情緒名で情緒推定する上では，人間による情緒推定と同水準で可能であることを実験により確認した．2008 年 8 月の 4 日ぶんについて情緒的表現の解析を実施したところ，約 124 万文から，情緒過程を表す箇所が 838,455 箇所あることが解析により分かった．これは非情緒過程を表す箇所(522,390 箇所)よりも多いことが分かった．ここからさらに，情緒対象が 73,706 種類抽出できた．情緒対象を情緒的偏りのある順に並べたところ，関心の持たれている事柄のキーワードが観測可能であることが確認できた．

残された問題としては，情緒対象となるキーワードに対する情緒名をより正確に解析することが挙げられる．今後，そのキーワードの使用されている文における情緒推定を追加することで，その解決を試みるべきであろう．また，本システムでは，用言を中心とした情緒推定であった．より広範囲に情緒を推定するために，副詞，接続詞，文末表現などの表現要素も情緒推定に取り入れる必要がある．

謝辞

本研究は，科学研究費補助金（若手研究（B）：19700149）の下で行いました．知識ベースおよびコーパスの開発に協力して下さった本学知能情報工学科池原研究室のメンバに感謝致します．

参考文献

[関根 05] 関根聡，武田善行，吉平健治：WEB 文書を対象とした KWIC システム， 自然言語処理， Vol.12， No.3， pp. 245-252 (2005)

[南野 04] 南野朋之，鈴木泰裕，藤木稔明，奥村学：blog の自動収集と監視， 人工知能学会論文誌， Vol.19， No.6, pp.511-520, (2004)

[Kanayama 08] Kanayama, H. and Nasukawa, T.: Textual Demand Analysis: Detection of Users’ Wants and Needs from Opinions, Proc. of Coling2008, pp.409-416 (2008)

[目良 02] 目良和也，市村匠，相澤輝昭，山下利之：語の好感度に基づく自然言語発話からの情緒生起手法， 人工知能学会論文誌, Vol.17, No.3, pp.186-195 (2002)

[池原 97] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎：日本語語彙大系，岩波書店，1997.

[田中 04] 田中努，徳久雅人，村上仁一，池原悟：結合価パターンへの情緒生起情報の付与， 言語処理学会第 10 回年次大会発表論文集, pp.345-348 (2004).

[黒住 06] 黒住亜紀子，村上雄弥，徳久雅人，村上仁一，池原悟：結合価パターン辞書における情緒表現性のある用言の意味分析，電子情報通信学会ソサイエティ大会講演論文集，基礎・境界，p.168 (2006)

表 5 情緒名ごとにみた情緒対象

順位	《P》の情緒対象	《N》の情緒対象	《A》の情緒対象	《S》の情緒対象
1	オンライントレード】	元気】	音楽	元気
2	情報集め	と	独占	事件
3	寝つき	判断	【フリーター	事故
4	少し情報量	【大卒フリーター】	話題	問題
5	広告費	【フリーター生活】	少し体調	男子生徒
6	豆乳	そう	【ライブドアFX	様々サイト
7	話し相手	気分	少しホットミルク	戦争
8	アルコール	あたり	僕	地震
9	広告効果測定システム	あなた	ビール	トラブル
10	ちょっとテレビ	マシン	毎日ぼーっ	破綻（はたん）危機
:				
19	バナナ	バイト】	プレゼン	現象
20	マカ	彼女	持ち	中毒
21	ダイエット日記	筋トレ	少し【メンズヴィトン長財布】	言葉
:				
39	最新情報	友達	業務	「新しく始まる枠（略）」
40	タバコ	タイミング	夫	中毒事件
41	【フリーター問題】	リスク	定職	アクション
:				
79	ファンタシースターポータブル	刃物男、男性	フォトグラファー	子
80	寒天	Web会議システム	ファン	高さ
81	口	友人	アルバイト	音
:				
159	業者然	一生	胸	インターネット物販、（略）
160	対立	チャンス	い	ホテル
161	県教委	料理	少し	感覚
:				
319	開会式	両左腕	生徒	雷がなり稲妻
320	ネットビジネス	同日午後	神様	雷、雨
321	ボーっ	手術法	魚町2番街	予想外
:				