

[DRAFT]

[DRAFT]

本記事は、情報処理学会誌「情報処理」(Vol.48, No.9, pp.995-1000, 2007)に掲載されているものの DRAFT 版です。

(ここに掲載した著作物の利用に関する注意)

本著作物の著作権は(社)情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。

[DRAFT]

[DRAFT]

# テキスト評価分析の技術とその応用

乾 孝 司<sup>†</sup> 奥 村 学<sup>††,†††</sup>

一般の人々の意見が含まれたテキストデータから、人々の何らかの評価に関する情報を抽出する技術を概観する。また、それら技術の利用の一例として、blog マイニングシステム「blogWatcher」の概要を紹介する。

## 1. Consumer Generated Media の登場

Consumer Generated Media (CGM) とは、ジャーナリストや新聞記者といった専門家ではなく、これまで情報の消費者となっていた一般の人々が、インターネットなどを通して自らコンテンツを生み出す新しいメディアの総称である。近年、爆発的な普及をみせている blog は CGM の代表例であると言えるが、他にも「@コスメ<sup>1</sup>」などの口コミサイト、「mixi<sup>2</sup>」などのソーシャル・ネットワーキング・サービスも CGM の一形態であると認識されている。

CGM は、一般の人々によってコンテンツが生成されるメディアであるため、既存メディアと比較して、表 1 中のテキスト例のような、個人の主観的な意見や個人の生活に密着した情報が含まれやすいという性質をもつ。

上記の性質と、CGM がもつコンテンツ量が増大傾向にあることが相まって、近年、CGM コンテンツを、新たな情報サービスの題材として、あるいは、マーケティング活動や企業のリスク管理活動における情報源として役立てようとする試みが行われ始めている。そしてこの結果、CGM コンテンツ、特に CGM に含まれる個人の意見に関する情報（意見情報）を分析、処理するための技術開発が盛んに行われるようになっていく。

## 2. テキスト評価分析

意見情報を分析、処理するための技術のうち、本稿ではテキスト評価分析<sup>3</sup>と呼ばれる技術を概観する。CGM に含まれる意見情報には、『安くして欲しい』と言った要望や、『安倍さんの意見に賛成です』と言った賛否態度など、いろいろな種類の情報がある。テキスト評価分析では、このような意見情報のうち、何らかの対象に関する評価の情報に注目する。

評価情報は、その評価内容に関して「良い/悪い」という観点から分類できる。本稿では、この観点のことを評価極性と呼び、「良い/悪い」のそれぞれの評価を、肯定的評価、否定的評価と呼ぶことにする。表 1 にそれぞれの評価極性をもつテキストの例を示す。テキスト評価分析における基本技術は、入力としてテキストが与えられた際に、そのテキストが肯定的評価を含むか、否定的評価を含むかを判定することであると言える。

テキスト評価分析の技術は既に産業活動等において活用されており、例えば「Yahoo! ブログ検索<sup>4</sup>」や、後述する「blogWatcher<sup>5</sup>」では、blog 中の評価（評判）情報を分析する機能を搭載した情報検索サービスやマイニングサービスが提供されている。

以下本稿では、まず、一般的な視点から、テキスト評価分析における基本的な処理の流れについて述べる。その後、テキスト評価分析の応用例として、blog マイニングシステム blogWatcher を紹介する。

<sup>†</sup> 東京工業大学 統合研究院

<sup>††</sup> 東京工業大学 精密工学研究所

<sup>†††</sup> (株) ブログウォッチャー

<sup>1</sup> <http://www.cosme.net/>

<sup>2</sup> <http://mixi.jp/>

<sup>3</sup> 英語圏では Sentiment Analysis と呼ばれることが多い。

<sup>4</sup> <http://blog-search.yahoo.co.jp/>

<sup>5</sup> <http://blogwatcher.pi.titech.ac.jp/>

表 1 評価極性をもつテキストの例

id	テキスト	評価極性
1	A社のスピーカ「BCD」は、とても心地よい音がでるので、お気に入りです。	肯定
2	運動した後だったので、ご飯もビールも美味しかった。	肯定
3	音が悪く、操作感もいまいちで、わずらわしい。	否定
4	先月「foo and bar」を購入しました。傷が付きやすいのが気になります。	否定
5	XYZ社の工場はいつもうるさい。	否定
6	鈴虫が鳴いている。	どちらでもない
7	自宅には電気ストーブと薪ストーブがあります。	どちらでもない

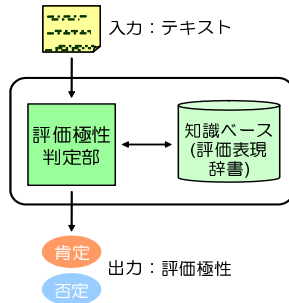


図 1 基本的な処理の流れ

### 3. テキスト評価分析の基本的な処理の流れ

図 1 に、入力テキストに対し、その評価極性を判定、出力する、最も基本的なテキスト評価分析の処理の流れを示す。評価（評判）情報の検索、マイニング等の応用システムは、図 1 に、検索、統計処理、可視化等の各モジュールを追加することで実現される。また、図 1 の入力テキストの粒度は、応用システムに応じ、フレーズや文、文書等、さまざまな単位に変化する。

図 1 において、入力テキストの評価極性は、評価極性判定部で判定される。この判定処理の際、評価極性判定部は、必要に応じて、必要な知識を知識ベースから獲得する。評価表現辞書は、知識ベースとして用意される知識の代表的なものである。評価表現辞書のエントリ例を表 2 に示す。この辞書には、評価を表す際に使われる単語（評価表現）と、それが表しやすい評価極性をペアとしたものが格納されている。

本章では、まず、評価表現辞書の作成手法の一例を紹介する。その後、評価極性判定部の処理の例として、評価表現辞書の知識を利用して文書の評価極性を判定する手法を説明する。説明する各手法は共に 2002 年に Turney<sup>4)</sup> によって提案されたものである。現在では、Turney の考えに基づいた多くの手法が開発されており、Turney の手法は、現在の基礎を築いた手

評価表現辞書の作成手法に関しては、実際の論文では、動的に語句の評価極性を推定する手法として提案されているが、本質は同じである。

法であると言える。

#### 3.1 評価表現辞書の作成

評価表現辞書は、単語や句に対し、それぞれが表しやすい評価極性を推定することによって作られる。ここでは便宜的に単語に限定して説明していくが、句の場合も同様である。

単語の極性推定問題を考えるにあたり、次のような仮定をおく。

仮定：

テキスト中において、肯定的（否定的）評価を表す記述の周辺には、やはり肯定的（否定的）評価が記述されやすい。

これは、テキストの局所的な文脈においては、評価極性は一致しやすいということを表している。この仮定は、語の視点から見れば「局所文脈内に現れる語は同じ評価極性を表しやすい」と捉え直すことができる。

上記仮定に従って、単語の評価極性を見積もる尺度を設計する。具体的には、典型的な評価表現との共起度という形で、上記仮定を尺度内に取り込む。

Turney<sup>4)</sup> では、まず、典型的な評価表現として excellent と poor を選ぶ。次に「altavista」検索サービスの NEAR operator を用いて典型表現との（10 単語以内での）共起情報を取得し、pointwise mutual information (PMI)<sup>1)</sup> によって共起度を測る。そして最終的に、これらの情報を統合し、式 (1) の尺度を提案した。

$$PMI(t, \text{excellent}) - PMI(t, \text{poor}) \quad (1)$$

評価表現	評価極性	評価表現	評価極性
良い	肯定	悪い	否定
優しい	肯定	脆い	否定
見易い	肯定	不味い	否定
美味しい	肯定	狭苦しい	否定
心地好い	肯定	煩わしい	否定

式 (1) で得られる値は評価極性値と呼ばれ、単語毎に異なる。単語  $t$  の評価極性値は、 $t$  が poor よりも excellent と共起しやすければ正に大きな値となり、逆の場合は負に大きな値をとる。これによって、評価極性値が正の値をとれば、 $t$  を肯定的評価を表しやすい語、逆に負の値をとれば、 $t$  を否定的評価を表しやすい語と推定することができる。応用によっては、「肯定 / 否定」の 2 値に落とさず、式 (1) の値をそのまま辞書に格納しておき、その値を利用する場合もある。また、実際に式 (1) を運用する際は、考慮する語  $t$  に対して品詞による制約を与えたり、低負荷で正確に PMI 値を求めるための工夫が必要になる<sup>4)</sup>。

### 3.2 レビュー分類

次に、評価極性判定部でおこなう処理の例として、評価表現辞書の知識を利用して文書の評価極性を判定する手法を紹介する。この極性判定課題では、文書として書評や映画評といったレビュー文書を扱うことが多いため、レビュー分類 (review classification) と呼ばれることが多く、本稿でも以下レビュー分類と呼ぶ。

レビュー分類の目的は、入力としてレビュー文書が与えられた時に、その文書が全体として肯定的評価を記述しているか、あるいは否定的評価を記述しているかを判定することである。レビュー分類をおこなうことによって、評価極性に従ってレビュー文書群を整理したり (図 2)、肯定的評価が記述されているレビュー文書だけを選択的に読むようなことが比較的容易に実現できる。

Turney が提案したレビュー分類のアルゴリズムを以下に示す。

- ステップ 1 評価表現辞書を参照することで、入力文書に含まれる評価表現を抽出する。この際、各評価表現に対応する評価極性値も同時に取得する。
- ステップ 2 ステップ 1 で得られた評価極性値の平均値を求める。
- ステップ 3 平均値の符号に応じて、入力文書の評価極性を判定する。もし、平均値が正の値であれば肯定的評価と判定し、そうでなければ否定的評価と判定する。

上記のアルゴリズムは、評価極性という観点で重み付き多数決をしていることに対応する。入力文書中の各評価表現は、それ自身が表しやすい評価極性側に式 (1) で与えられる重みつきで票を投じる。票の重みについては式 (1) 以外にもさまざまな設定方法を考え

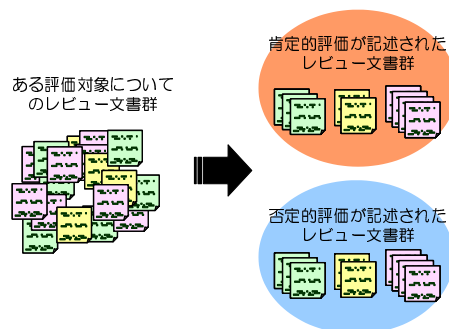


図 2 レビュー分類

ることができ、入力文書における評価表現の出現位置の情報を重みに反映させる場合もある<sup>3)</sup>。

入力文書に含まれる評価表現の数が極端に少なかったり、ステップ 2 で求めた評価極性の平均値が 0 に近い場合は、入力文書の評価極性を肯定的評価、否定的評価のどちらか一方に決定することは不相当となる。現実には、このような、評価分析という観点から見ればノイズと見なせる文書も多く存在するが、前記アルゴリズムでは対応できない。そのため、実際の運用では、ノイズ文書を適切に処理する工夫が必要になる。

## 4. 辞書利用 / 作成の注意点

テキスト評価分析では、評価表現辞書と呼ばれる知識ベースを用いる。そのため、テキスト評価分析の性能、あるいはその技術を利用した応用システムの性能は、評価表現辞書の質やサイズに依存する。本章では、評価表現辞書の利用段階、あるいは作成段階において、注意が必要な項目について整理する。

### 4.1 語形変化

日本語では、動詞や形容詞など、一部の語は語形変化を伴う活用語であり、当然、処理したいテキストデータ内では、活用した形で語が現れることがある。このため、仮に「美味しい」という評価表現が辞書に登録されていたとしても、表 1 の例 (id2) の「美味しかった」との辞書照合には単純には失敗してしまう。評価表現の中には形容詞も多く含まれるため、語形変化への対応は重要である。

この問題に対しては、あらかじめ考慮すべき活用形を展開して辞書に登録するか、辞書には基本形のみを登録し、照合の際、入力データ中の活用した語を基本形に戻して照合することで対応できる。基本形への変換には、例えば「JUMAN」や「茶筌」などの形

態素解析ツールが利用できる。

#### 4.2 表記の多様性

自然言語では、同じ意味の内容を複数の異なった表現を用いて伝えることができる。これは、テキスト上では「表記の多様性」として観察される。既存メディアでは、表記は規律的に統一されていることが多いが、非専門家である一般の人々によってコンテンツが生成される CGM では、この表記の多様性が生じやすい。

表記の多様さの種類には幾つかあるが、評価表現と関連の深いものとしては、字種違いの多様性がある。例えば、評価表現辞書に「煩わしい」という漢字仮名表記しか登録されていなければ、表 1 の例 (id3) の「わずらわしい」との照合には失敗してしまう。

この問題に対しては、複数表記を辞書登録しておくか、照合の際に、異表記ではあるが意味は等しいことを判定する機構を取り入れればよい。ただ、平仮名表記によって、意味の曖昧性が生じることもあるので、意味の曖昧性解消の処理も適宜必要になる。

#### 4.3 周辺文脈

前節で述べたレビュー分類手法は、評価表現の情報のみを考慮していた。しかし一般には、テキストに含まれる評価表現に加え、評価表現が現れる周辺文脈の情報を同時に扱った方が、より正確な判定ができる。扱いを慎重にしたい周辺文脈の例として、例えば「美味しくない」というフレーズでは、否定辞「ない」によって、評価極性が肯定から否定に変化する。また、「美味しいですか？」は評価ではなく単なる疑問（質問）の表明であるので、肯定でも否定でもない。

#### 4.4 評価情報と関連の強い語

評価表現以外にも肯定/否定の評価極性を推測することができる表現や語句は存在する。これらは評価表現ではないが、応用システムの用途に応じ辞書登録することで、評価分析の際の手がかりとして利用できる。

- 評価と関連して表出する態度や状態
  - － 嬉し涙（肯定）、冷や汗（否定）
- 特定の評価を導きやすい出来事や状態
  - － 表彰（肯定）、騒音（否定）

### 5. blogWatcher

本章では、テキスト評価分析の技術の応用例として、我々が現在開発中の blogWatcher について紹介する。

#### 5.1 システムの概要、機能

情報処理振興機構 (IPA) の 2003 年度の未踏ソフトウェア創造事業で採択頂き、blog の収集、監視、マイニングを行なうシステム blogWatcher の構築を我々は行ってきている。2004 年 8 月に一般公開を開

始し、2005 年 5 月には Ver.2.0 $\beta$  を、2006 年 5 月には Ver.3.0 $\beta$  を公開している (<http://blogwatcher.pi.titech.ac.jp>)。現在のシステムの主な機能は以下の通りである。

- キーワード検索
- キーワードの出現頻度推移表示
- burst 度の計算、表示 (トレンド分析)
- 評判情報検索 (評判分析)
- メタブログ
- ニュースとブログの自動対応づけ
- Versus Search
- blog の書き手の性別推定
- 行動分析

この内、評判分析については後で詳しく述べる。その前にここで、その他の幾つかの機能について簡単な説明を加えておく。

トレンド分析：あるキーワードが「いつ」「どの程度」注目されていたかについての情報を提供する。トレンド分析では、burst (キーワードの出現間隔が狭くなっている状態) を発見することによって、キーワードが注目されていると考えられる期間を検出する。あるキーワードが burst 状態であるかどうかは、blog ページ集合の中で、出現間隔が通常よりもどの程度短くなっているかを元にして計算されている。

メタブログ：システムが「今話題のトピック」と判定した話題を紹介するための blog (メタブログ) を自動生成する。システム内では、全単語について burst を計算することでその日に burst している単語がわかるため、その単語を元に注目されている話題を自動的に発見している。

Versus Search：入力されたキーワードに対して「ライバル」関係にあると考えられるキーワード候補を提示する。例えば「Windows」をキーワードとして入力すると、そのライバルとして「Linux」や「Macintosh」などが提示される。提示されたライバル候補を選択すると、最初に入力されたキーワードの分析結果と並べて、ライバルの分析結果を同時に閲覧することができる。

行動分析：blog 中の個人の「行動」を記述する個所を抽出し、その行動がどの時間帯 (朝、昼、夕方、夜) に行なわれたものかを自動推定する<sup>6)</sup>。blog には、個人の意見だけではなく、個人の日常の行動が記述されることが多く、それを分析対象とした初めての試みである。「ヨーグルトを食べるのは朝が多い?」といった疑問に答えられることを目指している。

## blogWatcher

評判情報検索

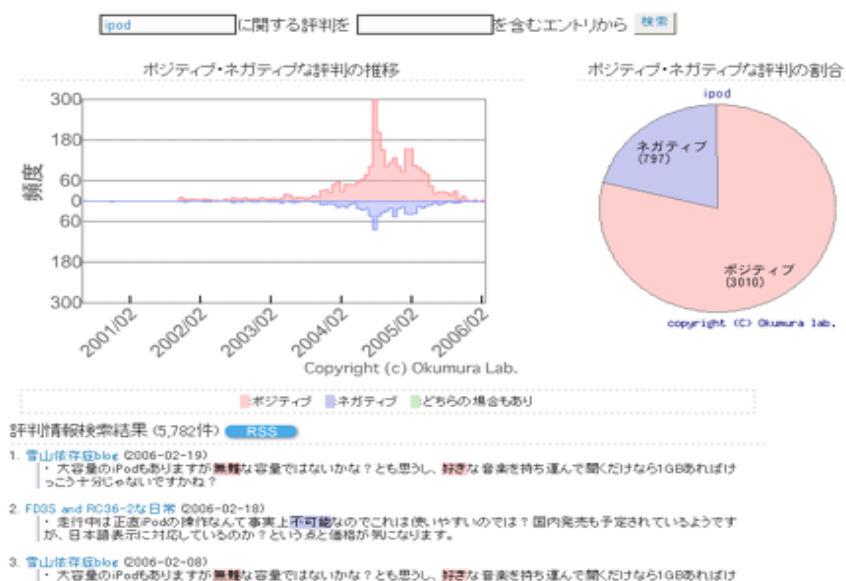


図3 blogWatcherによる評判分析のスクリーンショット(キーワード「iPod」で検索した場合)

## 5.2 blogWatcherにおける評判分析

blogWatcherでは、ある対象(キーワード)に関する評判情報を検索することができる。「iPod」をキーワードにして検索を行った結果を図3に示す。検索結果は、円グラフ、時間的推移を示すグラフおよび、キーワードに対して何らかの評価を行っている表現の一覧表示からなる。

評判分析における主な内部的処理は、blogから評価極性を含むテキスト断片を抽出することである。評価極性を含むテキスト断片を抽出するには、評価表現辞書を利用している。ただし、「あのHDDは容量が大きい」と「あのHDDは動作音が大きい」という2文における「大きい」が、肯定的評価であったり否定的評価であったりするように、何を評価しているかまで考慮して行わなければ肯定/否定を判定できない場合がある。そこでblogWatcherでは、評価対象や評価の際に着目する属性項目を考慮し、[HDD, 容量, 大きい]や[HDD, 動作音, 大きい]のように、

[評価対象, 対象の属性(着目点), 評価語]

の3要素の組を評価表現として定義し直した。そして、機械学習手法を用いることにより、約40万件の評価表現を自動獲得し、辞書情報として利用している。

評価表現(3つ組要素)の自動獲得手法として、我々は、少量の評価表現を手手でコーパスにタグ付けたタグ付きコーパスと、大量の生コーパス(タグを付し

ていないもの)を訓練データとして、semi-supervisedな機械学習手法を用いることで、少量の手手でタグ付けたコーパスから、大量の評価表現を自動獲得する手法を開発している<sup>2)</sup>。この手法では、タグ付きコーパスから自動的に評価表現のタグを付与する分類器を機械学習し、得られた分類器で生コーパスにタグを自動的に付与し、タグが付与されて新たに得られたタグ付きコーパスを機械学習のための訓練データとして追加した上で再度分類器を機械学習するステップを繰り返す。この繰り返しにより、少量の手手で付与したデータから大量の評価表現を自動的に得ることができる。

評価表現の評価極性を判定する際には、評価表現の3つの要素の情報だけでなく、評価表現が出現する文中の周辺の情報も利用できる場合がある。たとえば、文末に「(^-^)」など、「喜んでいる」顔文字がある場合、その文中の評価表現は肯定的であることが多いと考えられる。

また、評価表現が接続助詞などと相前後して出現する場合を考えてみる。

「軽くて便利」

「重いけど、デザインは素敵」

最初の例のように、順接の接続助詞を伴う場合、同じ評価極性の表現が連続して現れる傾向が強い。一方で、「けど」、「しかし」、「でも」等の、逆接の接続詞、接続助詞を伴う場合、相前後して出現する評

価表現は、極性が反転する傾向が強い。この性質を仮定するならば、あらかじめ少数でも肯定的あるいは否定的表現の辞書を持っていれば、テキスト中の、辞書に載っている表現の周辺に、肯定的あるいは否定的表現が存在するのではないかと推測できる。

評価表現が相前後して出現していなくても、この性質は全く同様である。たとえば、[ (PC の名称), , 重い] という否定的表現が辞書に登録されているとする。すると、この表現をテキスト中で探し、たとえば、

「 が重くなって、うんざり... 」

という文を見つけたとすると、「うんざり」が否定的な周辺情報であるということが推測できる。すると、今度は、この「うんざり」を手がかりに、テキストを探し、

「このハードディスクは音がうるさいので、  
うんざりだ」

という文が見つかったならば、[ハードディスク、音、うるさい] が否定的な評価表現の可能性があると推測できる。

## 6. blog マイニングのアプリケーション

我々の数少ない経験からではあるが、これまで述べてきたような blog を対象としたテキストマイニング技術は、少なくとも以下の 2 つの方面で期待感を持たれているようである。

- マーケティングリサーチ  
一般の人々の「生の声」を製品開発、企業活動に反映しようという、blog マイニングの典型的な応用と言える。マイニング結果は、レポートとして企業に提供される。
- WWW 上のポータルサイト  
「kizasi」に代表されるように、blog 上でのトレンドを可視化して見せる、また、口コミサイトと同様に、一般の人の評判を可視化して見せる等、一般のユーザに blog マイニングの結果を提供する。

我々の blogWatcher も、この原稿を書いている時点で公表できるものとしては、一部の機能が Yahoo! ブログ検索で使われているし、また、マーケティングリサーチでは、博報堂とサービスの実用化の準備中である。また、blogWatcher をベースとしたサービスとして、「体験検索 SHOOTI(シューティ)」が

(株) ブログウォッチャーにより 7 月 3 日に launch している。

## 7. 今後の課題：テキストの質の評価

私見ながら、今後重要となると考え、実際に我々も研究に着手しているのが、テキスト自体の評価である。言うまでもなく、blog のような個人が発信している大量の情報は玉石混交であり、それらの大量の情報を格付け、有用な情報のみをその中から選択する必要が出てきている。たとえば、spam という用語は、我々を悩まし社会問題化しているメール以外にも、今や blog、レビューなどについても使われるようになってきており、テキスト評価分析の際分析対象のテキストの質は無視できない要素と言える。分析する際に、優良なコンテンツのみを選別し、分析できればという要求があるのも事実である。今後は、テキスト評価分析自体の性能向上以外に、テキスト自体の評価手法の開発が重要であり、テキストの評価結果を反映した評価分析結果を提示できることが急務であると言える。

誌面の都合上、テキスト評価分析に関連するすべての事柄を紹介することは出来なかった。本稿の内容についてのより詳細な情報は、文献 5) から入手できるので、本稿と併せて参考にして頂きたい。

## 参 考 文 献

- 1) Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Proceedings of the 27th ACL*, pp.76-83 (1989).
- 2) Suzuki, Y., Takamura, H. and Okumura, M.: Application of Semi-supervised Learning to Evaluative Expression Classification, *Proceedings of the 7th CICLing* (2006).
- 3) Taboada, M. and Grieve, J.: Analyzing Appraisal Automatically, *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications* (2004).
- 4) Turney, P.D.: Thumbs up? thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *Proceedings of the 40th ACL*, pp.417-424 (2002).
- 5) 奥村 学, 大塚裕子, 乾 孝司: 意見分析エンジン-計算言語学と社会学の接点-, コロナ社 (2007).
- 6) 野呂太一, 乾 孝司, 高村大也, 奥村 学: テキスト中のイベントの生起時間帯判定, 情報処理学会論文誌, Vol.48, No.10 (2007).

属性の情報がないことを示す。

<http://kizasi.jp/>

<http://shooti.jp/>