Web 文書集合からの意見情報抽出と着眼点に基づく要約生成

立石 健二　福島 俊一[†]　　小林 のぞみ　上出 将行　高橋 哲朗
乾 孝司　藤田 篤　乾 健太郎　松本裕治[‡]

NEC インターネットシステム研究所[†]　　奈良先端科学技術大学院大学[‡]

1. はじめに

Web の意見は企業における市場調査・製品開発に重要な情報である。筆者らはこれまで、掲示板等の Web サイトから意見を抽出・分類する研究を進めてきた[1]。また、我々の研究と並行して、意見に着目した研究が行われている[4,5,6,7]。

これまでの研究の問題点として、意見の全体像を把握できるような要約機能が存在しないことがある。意見の要約は、商品間の比較を容易にするためや、分析の起点を得るために重要である。商品のための意見分析では、注目する商品と比較対象の同一な点と異なる点を明らかにした上で、異なる点を詳細に分析していくという手順をとるからである。

本稿では Web の意見をレーダーチャートの形式で要約する意見抽出分類方式を提案する。本方式では、意見を{対象物, 属性,評価}の 3 つ組で定義し、辞書と抽出ルールを用いた情報抽出のアプローチを採用して意見を抽出する。抽出した意見は着眼点と評価値の軸で分類し、レーダーチャートを作成できる。

２．機能要件

本研究の目的である Web の意見をレーダーチャートの形式で要約することを実現するためには次の機能が必要である。

(1) Web 文書から意見に該当する箇所を抽出する機能
(2) 抽出した意見を着眼点の軸で分類する機能
(3) 抽出した意見を評価値(肯定/否定)の軸へ分類する機能

これらの機能が実現できれば、図 3 のようなレーダーチャートの作成が可能である。図 3 の各軸は「安全性」「コスト」「走行性能」等の意見の着眼点が対応する。軸の値はその着眼点に属する意見全体の内の肯定意見の割合である。

３．アプローチ

意見を 2 つの条件を満たす情報と定義する。

- 表 1 の 3 つ組で構成される
- 記述者の判断として 3 つの構成要素間の関係が存在する

表 1　意見のモデル

種類	説明	例
対象物	商品,企業名,人名等	LaVie,NEC,アインシュタイン
属性表現	対象物の特徴・性質、対象物の部分	性能,価格,デザイン,サポート
評価表現	肯定又は否定の評価	良い,好き,速い,使いやすい

表 1 の 3 つ組は、従来の我々の対象物・評価の 2 つ組のモデル[1]を拡張したものである。この定義により意見を抽出する問題は、3 つ組を抽出する情報抽出の問題として扱うことが可能になる。

2 節の要件は、辞書と抽出ルールを用いた方式によって実現する。あらかじめ対象物、属性表現、評価表現辞書を用意しておき、3 つ組の抽出はこの 3 種類の辞書とこれらの表現間の関係を記述した抽出ルールによって行う。辞書の属性表現には、着眼点のラベルが、評価表現の辞書には評価値のラベルがあらかじめ付与しており、このラベルと手がかりとして意見を着眼点の軸、評価値の軸で分類する。

４．システム構成

本システムは、図 1 のように意見抽出部・意見分類部・辞書作成支援ツールで構成する。まず、Web 文書集合からあらかじめ用意した対象物・属性・評価表現辞書と抽出ルールを用いて意見を抽出する（4.2 節参照）。次に、属性・評価表現辞書に付与した分類ラベルを参照して意見を着眼点と評価値の軸で分類し(4.3 節参照)、レーダーチャートを作成する。

属性・評価表現辞書は、[2]で提案した方式を利用して対象物の分野毎に半自動的に作成する(4.1 節参照)。対象物名辞書はシステム利用時にユーザから与えるものとする。

本方式による意見の抽出分類イメージを図 1 の例を用いて説明する。図 1 では、入力として 2 つの文があたえられている。まず、システムこれらの文から{車 A,燃費, 良い} {車 B,加速, 鈍い}の 3 つ組を抽出する。それぞれの表現が対象物・属性・評価表現辞書に登録されていて、

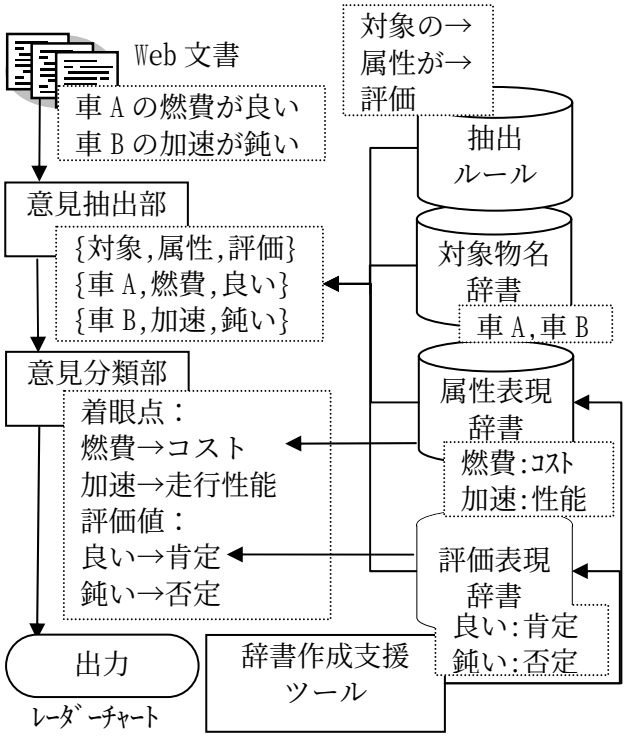

図 1　システムの構成

かつ 3 表現の組み合わせが当てはまる抽出ルールが存在する場合に、3 つ組を抽出する。次に、属性表現辞書に登録した「燃費」「加速」に対するラベルを用いて抽出した意見を「コスト」「走行性能」の着眼点に分類する。同様に評価表現辞書に登録した表現のラベルを用いて抽出意見を「肯定」「否定」の評価値に分類する。

４．１　属性・評価表現辞書作成

辞書作成支援ツールは、属性・評価表現辞書として小規模の初期辞書を用意しておけば、共起パタンを用いて大量文書集合からブートストラッピング的に双方の辞書を交互に増やすことができるツールである。ユーザは、ツールが提示した表現候補の採否を入力するだけで半自動的に大規模な辞書作成を効率的に実施できる[2]。

このような方式で収集する属性・評価表現に対して、着眼点・評価値の 2 種類のラベルを人手で付与する。着眼点の種類はユーザがその用途に合わせて自由に設定できる。例えば、車の分野では表 2 のような 9 つの着眼点が考えらる。一方、評価値の種類は肯定・否定・中立の 3 種類である。ラベルの付与は、原則としては、着眼点のラベルを属性表現(例、燃費：コスト，加速：走行性能)に、評価値のラベルを評価表現（例、良い：肯定、鈍い：否定）に付与する。ただし、共起する属性表現によって評価値が異なる評価表現が存在するため（例、価格が高い：否定，性能が高い：肯定）、それらには属性表現と評価表現の組み合わせに対して評価値を付与する。

表 2 作成した辞書の例

着眼点	属性表現の例
安全性(47)	ABS,エアバッグ,セキュリティ
コスト(52)	価格,維持費,燃費,修理代
サービス(35)	アフターサービス,営業,対応
走行性能(411)	エンジン,ギア,立ち上がり
耐久性(83)	ガタツキ,キズ,ボディ剛性
デザイン(155)	スタイル,外観,格好,形状
ブランド(147)	販売戦略,CM,ネームバリュー
メンテ(57)	アフターパーツ,オイル,交換
居住性(381)	エアコン,オーディオ,音,荷物
評価値	**評価表現の例**
肯定(728)	いい,エレガント,おしゃれ
中立(16)	普通,まあまあ,一長一短,並
否定(619)	イマイチ,嫌い,うるさい

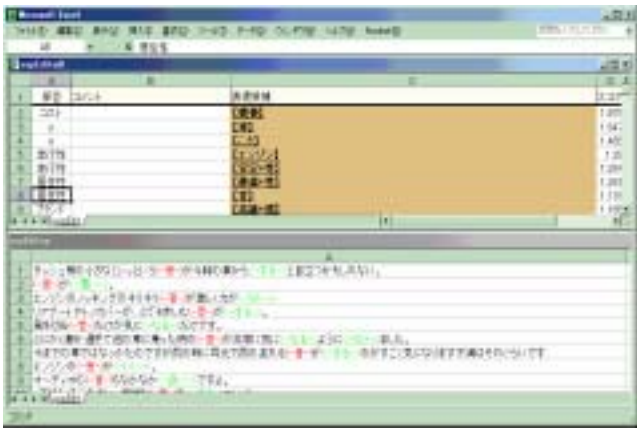

図 2　辞書作成支援ツールの画面例

図 2 に辞書作成支援ツールの画面例を示す。3 番目の列に属性表現の候補を表示しており、利用者は一番左側の列に表現の採否を入力する。表現を採用する場合は着眼点のラベルを、不採用の場合は”n”を入力する。表現の収集と同時にラベルを付与する。それぞれの属性表現の上をクリックすると例文が表示するため、実例を見ながら採否の判断が可能である。

４．２　意見抽出方式

意見抽出は 3 つの手順で進める。まず対象物・属性・評価表現の文書内の位置をそれぞれの辞書を参照して検出する。

次に、評価表現を中心として関係する属性表現・対象物を検出し、対象物・属性・評価の 3 つ組を抽出する。この処理は、属性→評価間の

表 3 抽出ルール(属性:属性表現，評価:評価表現，属性を対象物に置き換えたルールも用意)

抽出ルール	例
(属性:が|は|も|の|に|を|で|φ)→(評価)	デザインが良い,外観良い
(評価:<連体修飾>)→(属性)	良いデザイン
(属性)=(評価)	デザイングッド
(属性:の)→(*:が|は|も)→(評価)	デザインの質が良い
(属性:も|や|と|、)→([*]:も|は|が|で)→(評価)	デザインも広告も良い

関係と対象物→評価間の関係を検出する処理に分けて考える。どちらも、主語→述語の関係であり、同一の枠組みが適用可能だからである。また、表 3 の係り受け関係を利用した抽出ルールを用いる。抽出ルールは、辞書作成支援ツールで用いる共起パタンに類似するが、再現率を高めるため、より広い範囲の複数の文節に跨る関係も記述している。これらのパタンのいずれかに適合する場合は両者に関係があるとする。

最後に、抽出した 3 つ組の意見性を判定する。抽出した 3 つ組の中には「LaVie のデザインは良いでしょうか？（or 良いらしい。）」や「LaVie のキーボードが少し軽くなれば買いたい。」のように記述者の判断を示さないものも存在する。そこで、上記の抽出ルールとは別に評価表現の近傍の条件を示す接続詞、及び質問・伝聞を示す文末表現を考慮したルールを用意し、それらが存在する場合は意見でないとして除外する。

４．３　意見分類方式

原則として抽出した意見に付与した着眼点・評価値のラベルに従う。評価値の分類に関して、評価表現の近傍に奇数回の否定表現(例.ない)が存在する場合には評価値を反転する。

例）Mobile777 は良くない。→否定(奇数回)

例）Mobile777 は良くなくない。→肯定(偶数回)

５．分析例

4 節の方式を用いて作成したレーダーチャートの例を図 3 にしめす。このチャートは、車に関するレビューサイトの 630 記事から意見を抽出したものである。(a)のチャートは同車種間の比較であり、(b)のチャートは異車種間の比較である。(a)では、同車種であるためチャートの形状は類似していることがわかる。A1 より A2 の方がチャートの形状が全体的に小さく A1 の方が利用

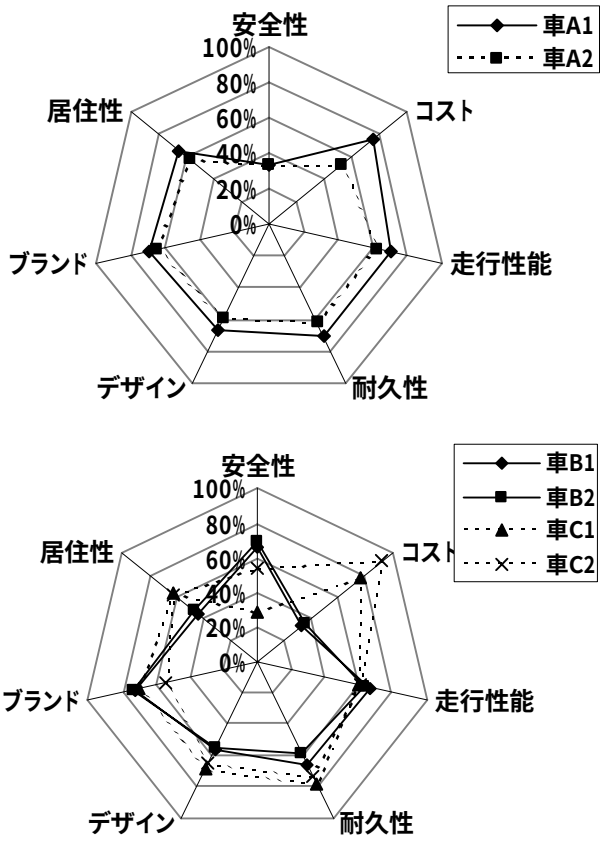

図 3　レーダーチャート作成の例

者の評価が高かったことがわかる。一方、(b)では 4 種類の車を比較しているが、B1 と B2、C1 と C2 は同系列の車であるためやはり類似したチャートの形状を持つ。B 系列の車は安全性と走行性能、ブランドイメージにおいて C 系列の車よりも評価が高い。C 系列は、コストが特に良く C2 においてはほぼ満点である。このように、意見をレーダーチャートという形式で要約することで商品の比較が容易になる。

次に、(a)のチャートの A1 と A2 の意見の相違点を詳しく分析する。表 4 は、走行性能に関する意見を出現頻度順に並べたものである。A2 は、「ナビ」「シフトショック」「装備」の否定意見が上位にあるため、ここが評価の差となったのではないかと読み取れる。このように、レーダーチャートによる要約を起点として、ポイントを絞った詳細な分析が可能となる。

チャートのために使用した辞書は、同サイトの 1000 記事から約 5 時間で作成できた。1445 表現の属性表現辞書と 1363 表現の評価表現辞書である。評価表現辞書のうち 53 表現は属性表現と評価表現の組に対して評価値を付与した。作成

表 4　着眼点「走行性能」に関する意見の例
（左:車 A1 右:車 A2）

属性表現	評価	数	属性表現	評価	数
エンジン	肯定	25	エンジン	肯定	21
トルク	肯定	13	運転	肯定	12
走り	肯定	12	走り	肯定	7
ﾊﾝﾄﾞﾘﾝｸﾞ	肯定	8	加速	肯定	7
パワー	肯定	8	ナビ	**否定**	6
加速	肯定	7	足回り	肯定	5
運転	肯定	7	パワー	肯定	5
走行性能	肯定	5	ｼﾌﾄｼｮｯｸ	**否定**	4
バランス	肯定	5	装備	肯定	4
装備	肯定	4	装備	**否定**	4
回転	肯定	4	ナビ	肯定	4

した辞書の例は、表 2 に示した通りである。属性表現に対しては 9 つの着眼点のラベルを用意したが、チャート作成に用いたのは意見の出現頻度が低かった 2 つを除いた図 3 の 7 つの着眼点である。構文解析ツールには CaboCha[3]を使用した。

６．関連研究

本研究では、情報抽出のアプローチを用いて、{対象物,属性,評価}の 3 つ組の意見を抽出分類しレーダーチャートという形式で意見を要約することを目的としている。

意見に注目した研究としては、(a)意見を肯定・否定に分類する研究[2]、(b)文を主観的な文と客観的な文に分類する研究[5]、及び、(c)意見に関連する表現を収集する研究[6][7]の 3 種類がある。(a)の研究は、レビューサイトの意見とそのレーティングを学習データとして、テキスト分類のアプローチで記事を肯定・否定に分類する。(b)の研究は、新聞記事の文を対象として、主観的な意見を示す文と、客観的な事実を示す文に分類する。(c)の研究は、[6]は主観的な表現を自動収集する方式、[7]は人の感情に関する表現を自動収集する方式を提案している。

しかし、(a)(b)(c)の研究ともに、{商品,属性評価}という意見の構成要素を取り扱うわけではないので、どの商品に対する意見であるか、どの着眼点に関する意見であるかまでを知ることはできない。そのため、レーダーチャートを作成するという研究目的達成には不十分である。

一方、情報抽出の研究領域では、意見の構成要素を抽出するという着眼点での研究はなかった。我々は今回、辞書と抽出ルールに基づく方式によって意見抽出を実現した。これは固有表現抽出の分野ではしばしば用いられる枠組みである。そのため、本研究はこれらで使われている技術を応用することが可能となる。

７．おわりに

本稿では、Web 文書から対象物・属性・評価の 3 つ組の意見を抽出・分類し、レーダーチャートを作成する方式を提案した。本方式では、情報抽出のアプローチを採用し、辞書と抽出ルールを用いて意見を抽出した。実際の分析例から、比較分析が容易になること、分析の起点にできることがわかった。

今後の予定として、以下のことを考えている。
(1) 意見抽出分類方式の評価：車以外の分野での本方式の有効性の検証、意見抽出・分類精度の評価を予定している。
(2) ルールの自動生成：4 節で用いた抽出ルールに関しては、直接係り受けはないが実質的に主語と述語の関係が存在する場合のルールの生成が問題となる（例、「良い点はデザインである。デザインが良くなる。」）。この件に関して、固有表現の要素間の関係を表すルールを自動生成する方式[8]を適用する予定である。
(3) 辞書作成の効率化：辞書作成支援ツールについては、ブートストラッピング的な手法で辞書を自動作成する方式[6]を応用して収集時間の短縮化を図る予定である。

参考文献

[1] 立石健二 石黒義英 福島俊一, "インターネットからの評判情報検索", 情報処理学会研究報告, NL-144-11, pp.75-82, 2001.
[2] 小林のぞみ 乾健太郎 松本裕治 立石健二 福島俊一, "テキストマイニングによる評価表現の収集", 情報処理学会研究報告, NL154-012, pp. 77-84, 2002.
[3] 工藤拓 松本裕治, “チャンキングの段階適用による日本語係り受け解析”, 情報処理学会論文誌, Vol.43, No.6, pp1834-1842, 2002.
[4] Kushal Dave, Steve Lawrence, and David M. Pennock, "Mining the peanut gallery: Opinion Extraction and Semantic Classification of Product Reviews", WWW2003, 2003.
[5] Hong Yu, and Vasileios Hatzivassiloglou, “Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences”, EMNLP2003, 2003.
[6] Riloff, E., and Wiebe, J., "Learning extraction patterns for subjective expressions", EMNLP2003, 2003.
[7] Liu, H., Lieberman, H., and Selker, T., "A Model of Textual Affect Sensing using Real-World Knowledge", IUI2003, 2003.
[8] Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman, “An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition”, ACL2003, pp. 224-231, 2003.