3F1-03

電子掲示板からの評価表現および評判情報の抽出

Extracting Evaluative Expressions and Reputations from the BBS

藤村　滋[*1]　豊田　正史[*2]　喜連川　優[*2]
Shigeru FUJIMURA　Masashi TOYODA　Masaru KITSUREGAWA

[*1]東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology,The University of Tokyo

[*2]東京大学生産技術研究所
Institute of Industrial Science,The University of Tokyo

Recently the power of Web communities is much attentioned. Many people check reputation of products which they want by BBS. So, companies cannot ignore reputation on the BBS. And they precisely check their product's reputation. Thus, automatically extraction of reputation and positive/negative classification has been getting more attention. This paper describes a reputation classifying method based on statistic extracting keywords. Some experimental results in classifying documents show the effect of the proposed method. Then, extracting opinions which include in reputation can be also realized by expansion of this method.

1. はじめに

買いたい商品について詳しく知りたいとき，掲示板でその評判を調べたことはないだろうか？しかし，目的の評判を探して掲示板を読む際，結果として大量のテキストを読むことになり，時間の浪費となってしまうことも少なくない．また，評判は掲示板のみでなく，レビューに関する記事，個人の日記やブログに書かれていることも多いが，そのような評判は従来の検索エンジンでは容易にはみつけられない．そこで，最終的にWeb 全体からの評判抽出につなげるために，我々が行った電子掲示板上からの評価表現および評判抽出について報告する．

2. 評判抽出の有効性と肯定・否定による分類

本報告では，対象に関する肯定・否定の意見を評判，および評判情報とする．評判抽出の有効性としては次のような例が挙げられる．

例えば，企業では新製品開発の際のマーケティングのために，自社製品の改善すべき点，他社製品の強みなどを知りたいという欲求がある．Web の特に個人サイトや掲示板の評判には，個人の意見がアンケート調査なしに手に入れられるという大きなメリットがある．また，クレーム処理の面から言えば，たとえ中傷であったとしても，自社の製品がネット上でどのように思われているかを知るのは非常に重要である．

個人の面においては，製品の購入時には我々は当然良い製品が欲しい．抽出された評判によって，悪評がすくなく，評価の高い製品を見つけるという意思決定支援が可能となる．

以上の点からも，評判抽出においては，抽出した評判が肯定的なのか否定的なのか分類したほうがより便利である．しかし，Web 上のページでは評判が分類されていることはめったにない．したがって，肯定的 (Positive)・否定的 (Negative) かどうかによる**文書分類 (document classification)** の必要性が生じる．

連絡先: 藤村滋，東京大学生産技術研究所喜連川研究室，
目黒区駒場 4-6-1，03-5452-6256，
fujimura@tkl.iis.u-tokyo.ac.jp

3. 関連研究

評判の抽出に関する先行研究としては，立石 [立石 01] らの研究があげられる．この研究では，ユーザが入力した商品名とあらかじめ辞書として用意した評価表現を近接演算する方法を用いて，インターネットの Web ページから意見を抽出している．また，抽出した意見の意見らしさ（適性値）を構文的な特徴を利用して判定している．しかし，この研究では評価表現辞書の作成，適正値判定処理どちらもヒューリスティックに構築されており，膨大な手間がかかってしまうことや，抽出可能な評判がかなり限定されてくること，さらには評価表現は話題のドメインによって大幅に変わるという問題点があった．

一方，掲示板のレビューを肯定・否定に分類し，抽出を行った例としては，Dave[Dave 03] らの研究がある．しかし，この研究での対象言語は英語のみであった．そこで，日本語でもこの手法が応用できそうかという点も含め，今回の報告ではこの論文の手法を参考にして実験を行った．この研究における手法の詳しい説明については，本報告で用いた手法も含め次章で報告する．

4. 本報告の手法

4.1 動機と処理の流れ

我々の最終的な目標は，Web 全体からの評判抽出を行うシステムを構築することにある．そこで，今回の報告では，評判が抽出できたと仮定して，その評判の肯定・否定分類[*1]に取り組むこととした．最終目標となるシステムに必要な機能については，図 1 に示す．本報告では図 1 の破線部分にあたる PN 分類を行い，その評判に対しどの程度肯定・否定の意味合いが強いかという観点で順位付けを行った．本手法の PN 分類部分では，次のことに注意した．まず，ドメインごとに限定されることがないようにする点である．次に分類器を容易に解析することができるという点である．後者については，評判を PN に分類する分類器は，評判の良し悪しを決定するルールそのものであるから，その分析を行うことによって，現在のトレンドや潜在的なニーズを掴むことができる可能性がある．

以上の要件を満たした PN 分類器作成のために，統計的に評

[*1] 以下，PN 分類と呼ぶ

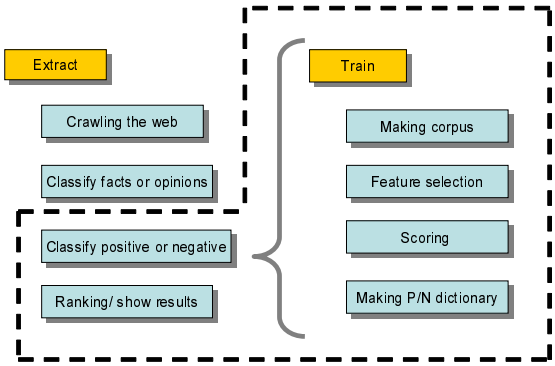

図 1: Process flowchart of our approach

価表現を取り出すことでドメイン依存の問題を解決した．コーパスから属性を選択し，評価表現としての重みをスコアリングし，評価表現辞書を作成する処理を行った．

4.2 処理の構成要素の説明

- **評価表現およびその辞書**

  本報告ではどのような意味で評価表現・評価表現辞書という言葉を用いるかについて次のように説明する．「評価表現とは，評判で用いられる特徴的な“ 語 ”のことであり，評価表現辞書とはその評価表現が肯定・否定どちらの表現であるかまで記された評価表現の集合である．」

- **訓練コーパス**

  今回の報告で実際に取り扱う評判の対象としては“ ノート PC ”を選んだ．評価表現辞書を統計的に作成するための訓練コーパスとしては，価格.com のノート PC に関する掲示板の 2003 年度の書き込みを用いることにした．肯定的な評判 935 件，否定的な評判 551 件である．価格.com では書き込みを行う人が，使用レポート（良），使用レポート（悪）のようなタグをつけることができ，これを利用することによってコーパス作成の手間を省くことができるというメリットがある．

- **属性の選択**

  評価表現の属性選択の手法としては，次の 2 種類の手法を試した．ひとつは (1) 形容詞，形容動詞のみを属性とする手法であり，もう一つは (2) (1)+名詞，未知語という手法である．

  形容詞，形容動詞については，主に日本語でモノの評価を表す表現であるので属性として採用した．名詞，未知語では「満足 (サ変接続名詞)」や PC のスペックなどが評価表現として取り込まれるのではないかという期待があったので採用した．

- **スコアリング手法**

  肯定的（否定的）な評判には，肯定的（否定的）な概念を持った語が多く含まれているはずである．この仮定を元に，肯定的な評判と否定的な評判の差をとる．一般的な語はどちらの文書にも同様に出現するはずであるから，その影響は打ち消される．評判において特徴的な語が肯定的な評価表現については正の値をもって，否定的な評価表現については負の値をもって抽出される．

表 1: Hi-scored features of adjectives, adj-verbs

Postive	明るい	bright	0.62
	綺麗	beautiful	0.60
	うれしい	happy	0.58
Negative	ひどい	awful	-0.60
	駄目	no way	-0.50
	不安定	not stable	-0.49

表 2: Hi-scored features of nouns,un-known words

Postive	満足	content	0.64
	SXGA	SXGA	0.58
	インチ	inch	0.57
Negative	修理	fix	-0.76
	最悪	worst	-0.68
	電源	power supply	-0.64

実際には，次のような式でスコアリングを行っている．

$$score(w_i) = \frac{P_P(w_i) - P_N(w_i)}{P_P(w_i) + P_N(w_i) + k} \quad (-1 \leq score(w_i) \leq 1) \tag{1}$$

ここで，$P_P(w_i)$ は肯定的な評判で属性 w_i が出現する確率である．同様に $P_N(w_i)$ は否定的な評判でのそれである．また k は，例えば $P_N(w_i)$ が 0 であった際に，$P_P(w_i)$ が 0.1 でも 0.8 でも結果としてスコアが 1 となってしまうという，1/1 の問題を解決するために分母に加えた実数である．

最後に，このスコアリングによって高いスコアを獲得した属性の例を，表 1，表 2 に示す．

5. 評価実験

5.1 PN 分類器としての性能評価

今回試した PN 分類法については，式 (2)，(3) に示す．

$$Score(d) = \sum_{ALL\,w_i} score(w_i) \tag{2}$$

$$\begin{cases} if & Score(d) > 0 \rightarrow positive \\ & Score(d) < 0 \rightarrow negative \end{cases} \tag{3}$$

各文書に含まれる属性のスコアの総和が 0 より大きければ，肯定的な評判であるとし，0 より小さければ，否定的な評判であるというように分類した．

分類の性能評価を行うため，比較対象として，C4.5 および SVM でも同様の実験を行った．C4.5 は決定木学習のアルゴリズムの一つであり，情報利得に基づいて分類規則を学習する．また，SVM は近年その高精度・高速性を理由に注目されている，パーセプトロン型の二値分類問題に対する機械学習手法である．SVM においては，ツールとして TinySVM[*2]を使用し，線形カーネルで実験を行った．他のオプションはデフォルトのままである．機械学習手法において与える属性については，スコアは用いずにその出現のみを考慮する形としたが，前章まで

*2 http://chasen.aist-nara.ac.jp/~taku/software/TinySVM/

表 3: Accuracy of P/N classification

	(1) adjs,adj-verbs		(2) (1)+n,UKW	
	Positive	Negative	Positive	Negative
Our approach	83.8	71.3	86.2	72.9
C4.5	79.1	60.5	78.0	60.3
SVM	79.6	71.8	80.4	73.0

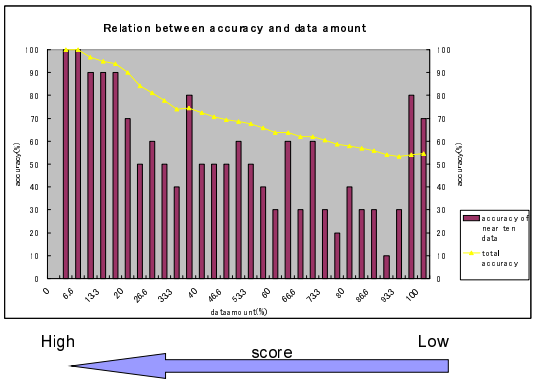

図 2: Relation between score and accuracy

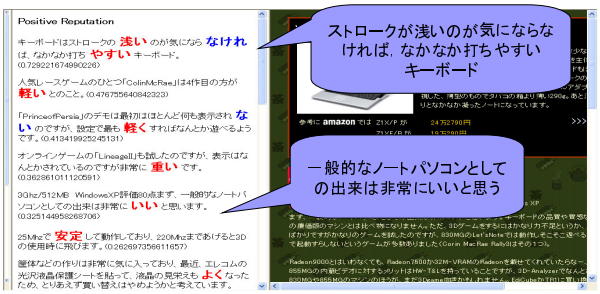

図 3: Example of positive reputation

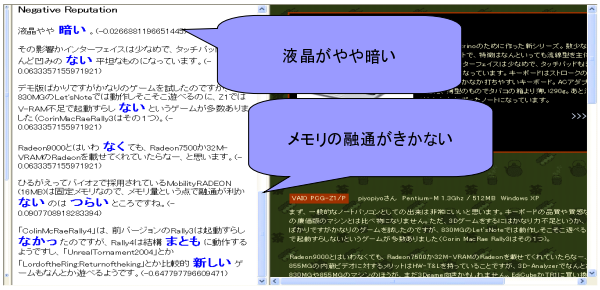

図 4: Example of negative reputation

で得られた属性と同様のものを用いた．訓練用のコーパスも同様に価格.com[*3]の 2003 年の肯定・否定の評判を用いた．

テストデータについては，価格.com の 2004 年[*4]の使用レポート（良）・（悪）の書き込み，それぞれ，240 件，137 件を評判として利用した．

各手法の分類精度については表 3 のようになった．本手法は C4.5 より P/N 分類に関して確実に精度が高く，SVM と比較をしてもまったく遜色のない分類精度が得られるという結果が得られた．

5.2 スコアと分類精度の関係

評判らしい文書を抽出するフィルタとして，この分類手法を応用できないかを検討するためにテストデータにわざとノイズをいれ，高スコアが得られた文章が評判そのものとなることを理想的な結果と想定し，実験を行った．

価格.com の掲示板 2004 年の書き込みに対し，評判とは異なる文書[*5]をノイズとして追加し，肯定的な評判，否定的な評判，ノイズを各 100 件になるようなデータセットを作成し，この実験でのテストデータとした．

この実験における結果は図 3 のようになった．

まず，スコアの絶対値が大きい順に分類されたデータを並べかえ，10 個を単位としてそこまでのデータ全体の精度を求めたものが図の折れ線である．また，付け加えた 10 個のデータの精度が図の棒グラフとなっている．図では，左から順に絶対値の大きい順にデータ量を増やしていき，一番右端では，折れ線はテストデータ 300 個全体での精度を表している．

この結果から，スコアが大きいものほど精度が高い，つまりスコアが大きいものは評判としても問題がないという結果が得られた．グラフの右端で直近 10 個のデータの精度が跳ね上がる傾向が見られる．これはノイズについては，評判でないというタグをつけて分類を行ったので，評判でないという意味で[*6]精度が高かったためにこのような結果になっている．

5.3 Web への拡張

最後に，今回の発表では，掲示板から得られた文書を中心に扱ってきたが，今後は実際の Web の文書へと拡張していきたい．そのための参考実験として，URL を入力するとその Web ページのテキストから評判を抽出してくる CGI を作成した．Google[*7]でクエリ「vaio　Z　レポート」で調べ，実際に Web ページを見て，評判が述べられているページ[*8]について，評判の抽出を行った結果について，図 3，図 4 に示す．

この結果から，抽出された文書についてはノート PC 自体というよりも，あるゲームをこの PC で遊ぶ際の状況についてのノイズも余分に抽出しているが，評判の抽出にも成功している．実際，この Web ページはかなりテキスト量が多い．ノート PC の評判を抽出できればその有効性は大きい．

6. 考察・検討

6.1 本手法の精度について

今回の実験の 1 つの目的であった，Dave らの手法を参考にした手法が，日本語においてどの程度有効であるかを調べるという点については，英語の場合では精度 85%程度であったのに対し，日本語ではおよそ精度 70%代後半であった．結果として精度については日本語のほうが 6%前後劣っていた．さらに，この実験の結果を得る前に，何度も予備実験を行った際，品詞による選別を行わないと精度が大幅に低くなることが得られた．英語では特にストップワードを設けるようなことをしなくとも問題がなかったのに対し，日本語では形容詞，形容動詞，名詞，未知語のように，評価表現としての属性を絞る必要性があることが得られた．

*3 http://www.kakaku.com/
*4 正確には 2004 年 4 月 8 日までの書き込みを使用した．
*5 例えば，ノート PC の使い方の質問であったり，特価情報の噂など
*6 スコアが 0.0 となっている
*7 http://www.google.com/
*8 http://www.kettya.com/notebook/sony/z.htm

また，機械学習による分類との精度比較では大きな差は得られず，SVM とだいたい同程度であることが示された．しかし，本手法には分類器の分析による知識の獲得という大きなメリットが存在する．SVM はその分類器は人間にとって可読不可能であり，なぜ入力文書が肯定的な評判であるのかは人間には理解不能である．また，C4.5 の決定木は人間にとって可読性はあるが，ある語が出現するかしないかだけの 1 方向の決定木になる傾向があり，そこから得られる知識は少ない．

実際の知識獲得の例については，次節で述べることとする．

6.2 P/N 評価表現辞書分析による知識獲得

この節では，分類器の一部である評価表現辞書を分析することで得られた知識について述べる．

表 1,2 に示した属性の例に着目すると，「明るい」「綺麗」「SXGA」「インチ」など主にノート PC では主に液晶について述べる際に用いられる語が高いスコアをもっている．ここから，ノート PC を購入する際には液晶に対する注目度が高いと推測される．また，「電源」という語自体がかなり否定的な評価表現であることを直感的には理解できない．しかし，このスコアは「電源」という語を用いた文書は大半が否定的な評判であるという事実を示す．

以上 2 つの推測を基に，コーパスの文書を実際に読んでみたところ，確かにノート PC の液晶に注目している人が多く，特に，ツルツルしたフィルムの様なものを張った液晶に対して注目度が高いことが得られた．また，「電源」については「電源が壊れる・故障する・入らない」といった類の文書が多く，確かに否定的な評判が多いことが得られた．

以上から，本手法では評価表現辞書を分析することによって，評判から一歩踏み込んださらなる知識を獲得できる可能性があることが示された．

6.3 精度とスコアの関係を調べる実験の際における属性について

精度とスコアの関係を確かめた実験の際，用いた属性は (1) 形容詞，形容動詞であった．(2) (1) ＋名詞，未知語の場合についても実験を行ったが，精度はデータ量が増えても振動するのみで期待された結果は全く得られなかった．

未知語は大半が名詞である．したがって，名詞を属性に加えることで実験結果が変わることとなる．この原因については，次のように解釈ができる．

前節で述べた「電源」の例について，確かに否定的な評判の中で使われる可能性が高い語であることは間違いない．しかし，例えば，「電源まわりがすばらしい」のように，肯定的な評判で使われる可能性も否定できない．つまり，形容詞，形容動詞に比べて，名詞は使われ方によって評判の良し悪しが変わってしまう可能性が大きい．評価表現辞書が統計的に作られているといっても，総計数十万単語からなるコーパスの中で多くても数十回程度しか使われない程度の属性がほとんどであるため，名詞の使われ方によって評判の良し悪しが変わる可能性は統計的な手法でも吸収しきれないほどに大きいと推測される．名詞を属性として採用する際には，形容詞よりも影響を弱くするように実数 $\alpha(\alpha < 1)$ をかけるなどの工夫が必要なのかもしれない．

7. まとめ・今後の課題

本報告では，日本語での評判の P/N 分類について，知識獲得が容易になるように統計的な処理を用いた手法について実装，評価実験を行った．本手法は従来から用いられてきた機械学習手法と比較してほぼ同程度の精度が得られることが分かった．また，高スコアの文書は評判そのものであることも確認し，Web 上から評判のような文書を抽出してくるような分類器として応用できる可能性があることを示した．最後に，今後 Web 文書への拡張を行っていくための参考実験として実際の Web ページから評判を取ってくるという例を示し，分類器から新たな知識を獲得できるという例についても示した．

以後，今後の課題について列挙する．

- **Web からの評判抽出システムの構築**
  今回の報告では，特に P/N 分類について注目し，その P/N 分類を評判自体の抽出にも役立てていけそうだという結果を導くことができた．今後は，Web 上の膨大な評判を集めるような crawler および評判と思われる意見の抽出を行う機能の構築を行い，掲示板に限らない，Web からの評判抽出システム全体の構築を行っていく．

- **さらなる精度の向上**
  名詞を属性に入れる際には形容詞・形容動詞と同格に扱うのではなく，何らかの工夫を行うことによってさらなる精度の向上が期待ができる．今後は例えば，名詞と用言の組み合わせまでも評価表現と考えるなど工夫を行っていく．

- **他ドメインへの拡張**
  本報告では，ノート PC に対象を絞って実験を行ったが，今後はデジカメや液晶ＴＶなどの他デジタル家電やレストラン，映画といったドメインの評判抽出にも拡張していく．

- **コーパス量と精度の関係の検討**
  上で述べたことと関連性があるのだが，評判のコーパスを作成したり，コーパスとして使えそうな文書を Web などから発見し，利用できる形に変換するのは容易な作業ではない．そこで，どの程度のコーパスの量があれば十分なのかを確認するために，コーパスの量と精度の関係について検討していく．

参考文献

[Dave 03] Kushal Dave,Steve Lawrence,David M.Pennock. Mining the Peanut Gallery:Opinion Extraction and Semantic Classification of Product Reviews. International World Wide Web Conference(WWW2003)pp.519-528,2003

[立石 01] 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.