系列パターンを利用した評価表現の分類

筬島郁子† 嶋田和孝‡ 遠藤 勉‡
† 九州工業大学大学院情報工学研究科 ‡ 九州工業大学情報工学部

1 はじめに

我々は，何らかの製品の購入を考えたり，どこかへの旅行を考えたりといった物事に対して判断を下す場合，第三者の意見などの評判情報を利用することが多い．WWW 上では，様々な人物との意見交換が可能であるため，我々は WWW を活用して第三者の意見などの評判情報の検索を気軽に行うことができる．しかしながら，WWW 上には大量の情報が溢れており，その収集した情報の内容を把握するには過大な労力を必要とする．企業においても，オンラインアンケートを実施するなどして，顧客からの評判情報の収集を試みており，近年，こうした評判情報を有効活用するためのニーズが高まってきている．

我々が評判情報を扱う場合，その情報が肯定的意見なのか否定的意見なのかを判断して参考にしている．よって，こうした第三者の意見などの評判情報を肯定的意見と否定的意見に自動分類することで，より情報の内容把握が容易になり，短時間で大量の情報を有効利用できる．そこで，本研究は，評判情報を文単位で扱い，評価文を肯定的意見と否定的意見に分類することを目的とする．ただし，評価文は抽出できたものと考え，その文が評価文かどうかはここでは議論しない．

2 関連研究

本研究では，文単位で評判情報を扱い，その文を肯定的意見・否定的意見に分類している．また，本研究は，系列パターンを利用したスコア計算に基づく手法により，評判情報を分類しているため，ここでは，Bag-og-words(以下，Bow) のみでのスコアリングにより評判情報を分類している関連研究について紹介する．

藤村ら [1] は，語のスコア計算をすることで，電子掲示板の評判情報を文単位で肯定と否定に分類している．分類する際の素性として，形容詞・形容動詞・名詞・未知語を利用している．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という仮定を基に，肯定的な評判と否定的な評判の差をとることで，スコアリング (式 2) をしている．$P_{pos}(w_i)$ は，肯定的な評判に語 w_i が出現する確率，$P_{neg}(w_i)$ は，否定的な評判に語 w_i が出現する確率を表しており，この確率 (probability) を利用した語 w_i のスコアを $Score_p(w_i)$ としている．また，$Score_p(s)$ は文 s のスコアを表している．

最終的な分類器は，文中の語のスコアリング (式 1) の総和 (式 2) を求め，その値が正の値ならば，肯定的な評価文，負の値ならば，否定的な評価文としている．このスコアリング手法を用いることで，分類結果は SVM と同等の精度が得られたとしている．

$$Score_p(w_i) = \frac{P_{pos}(w_i) - P_{neg}(w_i)}{P_{pos}(w_i) + P_{neg}(w_i) + k} \quad (1)$$
$$(-1 \leq score(w_i) \leq 1)$$

$$Score_p(s) = \sum_{ALL w_i \in s} Score_p(w_i) \quad (2)$$

3 肯定的意見・否定的意見の分類

3.1 事前処理

- **収集した評判情報の特徴**

  収集した評判情報は，ノート PC に関するレビュー記事で，一つは，記者が書いたレビュー記事[1]，もう一つは，一般ユーザが書いたレビュー記事[2]の 2 種類である．特徴として，記者が書いた記事は，一文が長く，重文で構成されていることが多く，また表現も固い．そのため，一般的に評価を表す形容詞 (「良い」「悪い」など) が少ない．また，記者が書いたレビュー記事は，否定的意見であっても，明確に否定している表現が少なく，一般ユーザが書いた記事に比べ，曖昧な表現が多いといった特徴がある．一方で，一般ユーザの書いた記事は，一文が短く，単文で構成されていることが多い．また，単純な形容詞などを利用した評価表現も多く，記者が書いたレビュー記事に比べると，肯定的意見と否定的意見の区別がつきやすい．

- **ラベル付けの方針**

  収集したレビュー記事において，評価文と思われる文に対して，人手で肯定的意見 (Positive)，否定的意見 (Negative) のラベル付けを行う．前述した通り，記者が書いた記事には，重文が多く存在するため，一文であっても，肯定的意見と否定的意見が混在することが多い．よって，より有効な素性を抽出するために，表層的な手掛かり語を用いて，逆接の意味をなす文のみ単文に分割 (表 1) し，ラベル付けを行う．しかしながら，分割処理適用後も，肯定と否定の評価が混在している文も存在する．このような文に対しては，結論として，その文がどちらの内容であるかによって判断する．

3.2 素性選択

- **系列パターンの利用**

  評価文を肯定的意見と否定的意見に分類する際，素

[1] www.sbpnet.jp/vwalker/review/index_c.asp?page=1
[2] http://review.japancnet.com

表 1: 逆接の分割処理の例

手掛かり語	分割処理
が，	使い勝手はよい**が，**／ 大きすぎて持ち運びにくい.
だが，	コンパクトで運びに便利**だが，**／ 連続駆動時間が短い.

性選択は重要である．単なる Bow のみの素性では，人手であってもその素性が肯定的意見のものなのか，否定的意見のものなのかを判断するのは困難である．例えば，PC 製品において,「CPU」「HDD」「高い」といった名詞や形容詞のみではその素性が肯定もしくは否定のどちらの意味を表しているのか判断するのは難しい．しかしながら,「処理＋能力＋高い」「PC ＋値段＋高い」といった語の組合わせでみてみると，その判別が容易になると期待できる．そこで，今回は素性として，Bow だけでなく，系列パターンも用いることとする．その結果，Bow だけでは，どちらのクラス (肯定的意見・否定的意見) に属するのか判断できない評価文であっても，系列パターンでの判断を加えることで分類できる．また，系列パターンの抽出には，PrefixSpan [3] アルゴリズムを用い，その実装には，工藤氏が開発したソフトウェア[1]を用いた．

表 2: 素性として利用した品詞

品詞の種類
名詞，形容詞，形容動詞，動詞 (サ変動詞を除く)，可能を表す助動詞 (〜デキル)，否定を表す助動詞 (〜ヌ，〜ズ)，未知語

評判情報を扱った関連研究のほとんどは，分類に利用する素性に対して，品詞の制限を行っている．本研究においては，系列パターンを抽出する際，品詞の制限を厳しくすると，ノイズとなる系列パターンは除外できるが，抽出できるパターン数も減少してしまうという問題がある．そこで，より多くの系列パターンの抽出のため，少しでも分類に有効だと考えられる品詞 (表 2) は利用する．

- **低頻度の系列パターンの利用**

テキストマイニングや重要語抽出など様々な分野において，高頻度語が最も信頼できる語であると判断するのが一般的である．系列パターンにおいても，パターン長が長い程，文に近いパターンが抽出でき，minimum support 数 (最小頻度数：以下，ms 数) が高い程，他の文に多く出現することから，パターン長および ms 数は高い方が信頼できる．しかしながら，現実問題として，人手で書かれた評価文からは，大量のデータからでないと，ms 数の高い系列パターンは抽出しにくい．その要因として，人手で書かれた評価文には，表記のゆれが多いことがあげられる．例えば，ある人物は，肯定的な評価をする際に「よい」という形容詞を用い，ある人物は「いい」と答える．表層だけみると，これらは全く別の語となってしまう．こういった表記のゆれにより，低頻度の系列パターンの中にも，分類に有効な素性が多数存在することとなる．そこで，ms 数が高い系列パターンのみを素性として利用するだけでは，素性の数が少なすぎるという問題を解決するために，有効な低頻度の系列パターンも素性として利用する．

しかしながら，低頻度の系列パターンにはこうした素性として有効な系列パターンも存在するが，同時に，ノイズとなる素性も多く存在する．よって，有効な系列パターンが含まれるという理由だけで，低頻度の系列パターンすべてを加えるのは強引すぎる．そこで，低頻度の系列パターンの中で素性として有効な系列パターンのみを底上げするための尺度が必要となってくる．この低頻度の系列パターンを底上げするための尺度として利用するものを以下にあげる．

尺度 1：IDF

肯定的意見か否定的意見かに分類する際，片方のクラスに偏って出現する語は，そのクラスを表す重要な手掛かりとなるという考えの基，式 3 より算出する．$pos(w_i)$ は，語 w_i の肯定的評判に含まれる数を表し，同様に $neg(w_i)$ は，語 w_i の否定的評判に含まれる数を表している．また，$\sum pos$ と $\sum neg$ は，肯定的評判に含まれる語の総数と否定的評判に含まれる語の総数である．語 w_i の肯定的評判での IDF 値を $IDF_{pos}(w_i)$ で，語 w_i の否定的評判での IDF 値を $IDF_{neg}(w_i)$ で表す．

$$\begin{cases} IDF_{pos}(w_i) = \log\left(\dfrac{pos(w_i)+1}{\sum pos} \times \dfrac{\sum neg}{neg(w_i)+1}\right) \\ IDF_{neg}(w_i) = \log\left(\dfrac{neg(w_i)+1}{\sum neg} \times \dfrac{\sum pos}{pos(w_i)+1}\right) \end{cases} \quad (3)$$

尺度 2：片方のクラスにのみ出現する語

肯定的意見か否定的意見かに分類する際，片方のクラスにのみ出現する語は，そのクラスを表す重要な手掛かりとなるという考えの基，尺度として用いる．

低頻度の系列パターンの底上げには，低頻度の系列パターンのうち，各クラスの IDF 値 (式 3) 上位の語かつもう一方のクラスでの IDF 値上位でない語を含む系列パターン，もしくは低頻度の系列パターンのう

[1] http://chasen.org/~taku/software/prefixspan/

ち，片方のクラスにのみ出現する語を含む系列パターンを用いる．

3.3 分類器

提案する分類器は，スコア計算に基づく手法で，まず Bow のみを用いたスコア計算を行い，その後，Bow のみでは分類が難しい文に対して，系列パターンを用いたスコア計算を行う．Bow のスコア計算は，藤村ら [1] が提案している確率を利用する手法 (式 2) と IDF 値 (式 3) を利用する手法の 2 種類である．IDF 値を利用する手法とは，文中に出現する語すべてにおいて，肯定的な評判での IDF 値と否定的な評判での IDF 値を算出し，それぞれのクラスにおける IDF 値の総和の差分 (式 4) をその文 s のスコア $Score_{IDF}(s)$ とする方法である．

$$Score_{IDF}(s) = \sum_{ALL w_i \in s} IDF_{pos}(w_i) - \sum_{ALL w_i \in s} IDF_{neg}(w_i) \tag{4}$$

まず，文 s のスコア $Score(s)$ を，$Score_p(s)$ もしくは，$Score_{IDF}(s)$ で評価する．$Score(s)$ が正の値ならば，肯定の評価文とし，負の値ならば，否定の評価文とする (式 6)．

$$Score(s) = \begin{cases} Score_p(s) & \cdots(\text{式 2 参照}) \\ Score_{IDF}(s) & \cdots(\text{式 4 参照}) \end{cases} \tag{5}$$

$$\text{if} \begin{cases} Score(s) > 0 \rightarrow Positive \\ Score(s) < 0 \rightarrow Negative \end{cases} \tag{6}$$

しかしながら，$Score(s)$ が 0 に近い場合は，どちらのクラスに属する評価文か曖昧である．そこで，スコア値の絶対値が 1 より小さい場合は，Bow のみのスコア計算では不十分であることから，系列パターンを利用したスコア計算 $Score_{Pat}$ を適用する (式 7)．

$$Score(s) = Score_{Pat}(s) \quad | Score(s) | < 1 \text{ のとき} \tag{7}$$

系列パターン Pat のスコア計算は，訓練データ中の肯定の系列パターン $Learn_{pos}$ のうち，文中に存在する系列パターン数 $num(Pat_{pos}(s))$ と，訓練データ中の否定の系列パターン $Learn_{neg}$ のうち，文中に存在する系列パターン数 $num(Pat_{neg}(s))$ の差分 (式 10) を利用する．また，低頻度の系列パターンを底上げする場合は，その底上げする系列パターンも訓練データ中の高頻度の系列パターンと同等に扱い，系列パターンのスコアを算出する．

$$Pat_{pos}(s) = \{Pat | Pat \in Learn_{pos} \wedge Pat \in s\} \tag{8}$$

$$Pat_{neg}(s) = \{Pat | Pat \in Learn_{neg} \wedge Pat \in s\} \tag{9}$$

$$Score_{Pat}(s) = num(Pat_{pos}(s)) - num(Pat_{neg}(s)) \tag{10}$$

最終的な文 s のスコア $Score(s)$ は，同様に式 6 で分類する．

4 分類実験と考察

分類実験は，10 分割交差検定 (訓練データ：9/10，テストデータ：1/10) により行った．実験に使用したデータの内訳は以下 (表 3) の通りである．精度の比較として，藤村らの手法をあげているが，素性の品詞は，予備実験において，本手法の品詞制限が良好であったため，本手法と同じ品詞制限とする．評価には，適合率 (P)，再現率 (R)，および F 値 (F) を利用する (式 11，式 12，式 13)．実験に利用した素性とスコア計算式は，表 4 の通りである．

表 3: 実験データ内訳

記事	文の内訳
(A) 記者が書いたレビュー記事	肯定：393 文
(115 記事)	否定： 95 文
(B) 一般ユーザが書いたレビュー記事	肯定：745 文
(69 記事)	否定：507 文

$$P = \frac{\text{正解した文数}}{\text{そのクラスに割り当てた文数}} \tag{11}$$

$$R = \frac{\text{正解した文数}}{\text{そのクラスに割り当てるべき文数}} \tag{12}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{13}$$

訓練データより，素性とする系列パターンを抽出する際，系列パターン長と ms 数のバランスが重要となってくる．系列パターン長と ms 数を変化させながらの予備実験を行った結果，最も精度の良かった組合せは，2 種類ともパターン長が 2-3 で，ms 数においては，肯定が 4 で否定が 2 であった．また，この組合せの系列パターンにおいて，ベースとする Bow のスコア計算式の違いによる精度変化を確認するための実験結果を以下 (表 5，表 6) に示す．

表 4: 分類実験に利用した素性とスコア計算式

素性とスコア計算式
(1) $Score_p$(式 2)
(2) $Score_{IDF}$(式 4)
(3) 系列パターン
(4) (3) ＋尺度 1 で底上げした低頻度の系列パターン
(5) (3) ＋尺度 2 で底上げした低頻度の系列パターン
(6) (3) ＋すべての低頻度の系列パターン

実験結果 (表 5) から，記者が書いた記事においては，Bow のスコア計算式は，式 2 の方が良い精度結果を得た．記者が書いた評価文は，IDF 値による分類結果が悪いことから，意外にも肯定と否定の文に特徴だった語が存在しないといえる．これは，記者が書いたレビュー記事においては，否定の表現をあまり使わないという特徴が要因と考えられる．

ユーザが書いた記事においては，実験結果 (表 6) から，適切な Bow のスコア計算式は，肯定と否定で異なっていた．しかしながら，肯定と否定の精度の全体的なバランスをみた場合，式 4 の方が良い．また，評価文を収集する際にも，否定の表現は肯定の表現に比べて少ないために，否

表 5: Bow のスコア計算式の違いによる分類 (A)

スコア計算式	素性			精度		
	クラス	length	ms 数	P(%)	R(%)	F
(1)	Positive	2-3	4	90.94	89.48	**90.20**
	Negative	2-3	2	60.59	61.87	**61.22**
(2)	Positive	2-3	4	97.87	47.24	61.89
	Negative	2-3	2	29.39	95.49	45.85

表 6: Bow のスコア計算式の違いによる分類 (B)

スコア計算式	素性			精度		
	クラス	length	ms 数	P(%)	R(%)	F
(1)	Positive	2-3	4	73.97	69.73	**71.79**
	Negative	2-3	2	60.52	65.32	62.83
(2)	Positive	2-3	4	78.48	60.85	68.54
	Negative	2-3	2	57.99	76.25	**65.87**

※ length … 系列パターン長

定の評価文は抽出しにくい傾向があることから，否定の精度が良い式 4 の方が適切である．

実験では，素性として妥当であった系列パターン (パターン長：2-3，ms 数 肯定：4 否定：2) に対して，さらに底上げの尺度を用いて選別した低頻度の系列パターンを加えた場合，底上げの尺度を利用せずに低頻度の系列パターンすべてを加えた場合の比較，およびに手法の違いによる精度比較として，Bow のみを素性に用いた藤村らの分類と SVM 学習器[3]を用いた分類，本手法による分類の 3 つの手法を比較した (表 7，表 8)．SVM の素性には，系列パターンと片方のクラスにのみ出現する語を利用し，素性に対する重みの付与，パラメータの変更は行っていない．

記者が書いた記事 (表 7) においては，尺度 1 を利用して，低頻度の系列パターンを底上げしたものを加えた分類結果が最も良い精度であった．これは，低頻度の系列パターンであっても分類に有効な系列パターンが存在することを表している．また，尺度 2 を利用した分類では，精度の低下がみられたため，この尺度は有効でないといえる．その要因として，この条件の場合，偶然にも片方のクラスにのみ一回だけ出現した名詞も含まれてしまうので，本来，肯定と否定の分類に適さない系列パターンも底上げしてしまったことが考えられる．

一般ユーザが書いた記事 (表 8) においても，記者が書いた記事と同様に，尺度 1 を利用して，低頻度の系列パターンを底上げしたものを加えた分類結果が最も良い精度であったが，肯定の文においては，低頻度の系列パターンをすべて加えた場合に，若干の精度向上がみられる．これは，低頻度の系列パターンの中に，まだ，有効な素性が残っていると考えられるため，他の尺度を検討することにより，さらなる精度向上が期待できる．

二種類の記事において，実験結果から，低頻度の系列パターンの中にも素性として有効な系列パターンが存在することを確認できた．また，手法の違いによる比較では，SVM での分類において，一見，否定的意見での Precision が高いようにみえるが，これは，Recall がかなり低いことから，あまり信頼性のある精度とはいえない．また，ユーザが書いた記事では，SVM での肯定的意見の分類精度が最も良いが，これも否定的意見の分類精度が悪いことから，全体的なバランスは良くない．よって，手法ごとの精度比較においても，本手法が最も良い結果を得ることができたといえる．

[3] http://svmlight.joachims.org/

表 7: 低頻度の系列パターンを利用した分類 (A)

素性	Positive			Negative		
	P(%)	R(%)	F	P(%)	R(%)	F
(1)+(3)	90.94	89.48	90.20	60.59	61.87	61.22
(1)+(4)	90.97	90.19	**90.58**	61.74	61.87	**61.80**
(1)+(5)	90.64	88.40	89.51	56.02	61.04	58.04
(1)+(6)	91.01	87.87	89.42	56.41	63.24	59.63
藤村ら	91.10	88.56	89.91	57.47	63.35	60.27
SVM	83.17	99.09	90.44	64.00	23.15	34.00

表 8: 低頻度の系列パターンを利用した分類 (B)

素性	Positive			Negative		
	P(%)	R(%)	F	P(%)	R(%)	F
(2)+(3)	78.48	60.85	68.54	57.99	76.25	65.87
(2)+(4)	78.62	60.85	68.60	58.04	76.43	**65.98**
(2)+(5)	66.29	69.09	67.66	53.67	50.43	52.00
(2)+(6)	78.22	61.54	68.89	58.26	75.68	65.83
藤村ら	77.07	60.94	68.06	57.49	74.31	64.83
SVM	65.14	68.79	**74.41**	64.46	33.92	44.55

5 おわりに

系列パターンと Bow を素性としたスコア計算により，評判情報を肯定的意見と否定的意見に分類した．実験結果から，評判情報の分類において，系列パターンを用いることで，Bow のみで分類した結果よりも，若干ではあるが，精度の向上がみられ，系列パターンは，評価文の分類に有効な素性としての潜在的な素質があることを確認できた．

今後，大量のデータでの実験により，素性としての系列パターンの有効性の検討や訓練データの肯定と否定の割合による精度変化を確認する必要がある．

参考文献

[1] 藤村 滋，豊田 正史，喜連川 優: 電子掲示板からの評価表現および評判情報の抽出，人工知能学会全国大会 (第 18 回)，2004

[2] 山崎 貴宏，新保 仁，松本 裕治: 系列パターンを素性とした論文概要文の自動分類，人工知能学会「人工知能基礎論」,「知識ベースシステム」合同研究会，信学技報 AI2002-83，pp.13-18，March 2003.

[3] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, M. Hsu : PrefixSpna :Mining Sequential Efficiently by Prefix-Projected Pattern Growth, Proc. International Conference of Data Engineeringm (ICDE), pp.215-224 (2001).