スコアリングを基にした評価文の分類

河野勇介† 嶋田和孝‡ 遠藤 勉‡

†九州工業大学大学院情報工学研究科情報科学専攻
‡九州工業大学情報工学部知能情報工学科

概要

本論文では，Web 上の掲示板などの文章から収集した意見情報や評判情報を含む文を肯定的意見と否定的意見に分類することで，情報の内容把握をより容易にすることを目指す．提案する分類器は，単語の出現情報を tf・idf の概念や品詞による重みを考慮した式によってスコア化し，そのスコアを手がかりにして文を分類する．実験により，提案手法が SVM など他の手法よりも高い精度を得られることを実証した．また，ラベルなしデータから自動的に学習データを拡張させる手法についても紹介する．この手法は元の学習データが少ない場合に有効であるという結果を得た．最後に，分類した評判情報の活用例を紹介する．

Sentence Polarity Classification Based on Scoring Method

Yusuke KAWANO†, Kazutaka SHIMADA‡, and Tsutomu ENDO‡

†Graduate Scool of Computer Science and Systems Engineering, Kyushu Institute of Technology
‡Department of Artificial Intelligence, Kyushu Institute of Technology

Abstract

In this paper, we propose a method for classification of sentences containing a user's opinion, such as positive and negative. Identification of the polarity leads to a benefit in terms of readability for readers of Web documents. Our method computes a score based on the word frequency. It is an extension of $tf{\cdot}idf$ method. Also we apply weighted part of speech tags to the scoring process. We classify sentences into positive and negative by using the score. First, we compare our method with other methods, such as SVMs. In the experiment, our method obtained the best performance. Next, we consider the effectiveness of an approach using unlabeled data. Finally, we explain some applications with our method.

1　はじめに

我々は，何らかの製品の購入を考えたり，旅行の行き先を考える場合，第三者の意見を参考にすることが多い．このような情報を得るためには，以前は，直接お店に行って調べたり，身近な人に尋ねたりすることしかできなかったが，Web の普及に伴い，現在では，自宅のパソコン等を利用して不特定多数の人から大量の情報を得ることができるようになった．Web 上に存在するこのような数多くの評判や意見情報は，近年，企業などでも重要視されている．一般ユーザから寄せられた意見情報などは，新しい製品の開発に活かすことができる．しかしながら，人手で Web 上に溢れる大量の情報を収集し，その内容を把握するには過大な労力を必要とする．そのため，このような情報を自動的に収集し，効率良く利用するための様々な研究が盛んにおこなわれている [1].

我々は Web 上の掲示板などの文章から収集した意見情報や評判情報を含む文を肯定的意見と否定的意見に分類することで，情報の内容把握をより容易にすることを目指している．本稿では，スコアリングを基にした分類器を紹介し，この分類器によって分類したデータを利用した情報活用の例を紹介する．

2　関連研究

本研究では文単位で肯定的・否定的に分類しているが，ここでは，何を単位として分類しているかは問わず，評判情報を分類している関連研究について紹介する．

2.1　Pang らの手法

ここでは，評価文書分類に初めて機械学習を適用した Pang ら [2] の研究を紹介する．Pang らは，教師あり機械学習による分類手法が，評価文書分類において有効であるかを検証した．彼らはいくつかの学習器で分類実験をおこない，その中でサポートベクターマシン (SVM) が最も良い精度が得られたとしている．SVM は，訓練データを正例と負例に分離し，かつ，正例と負例のマージンが最大になるような超平面を求める学習器である．

2.2　藤村らの手法

我々の研究と同様にスコアリングを用いた手法で評価文書を分類している研究に藤村ら [3] の研究がある．藤村らは，日本語で書かれた掲示板のレビューを肯定・否定に分類する研究をおこなった．彼らの研究では，肯定的な評判には肯定的な概念，否定的な評判には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，肯定的な評判で単語 w_i が出現する確率と否定的な評判で単語 w_i が出現する確率の差をとっている．実際には，次のような式でスコアリングをおこなっている．

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

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．k は 1/1 の問題を解決するために分母に加えた実数である．最終的な分類は，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とし，0 より小さければ否定的というように分類している．藤村らはこの手法を用いることで，SVM と同等の分類精度が得られたとしている．また，彼らの研究は文書レベルでの分類であるが，そのまま文レベルの分類にも適用できる．

3　スコアリングを基にした分類器

ここでは，スコアリングを基にした分類器を用いて，意見情報や評判情報を含む文を肯定的意見と否定的意見に分類する具体的な方法について述べる．

3.1　分類器

本研究では，文に含まれる単語そのものの情報によって分類を試みる．具体的には，まず，形態素解析ツール茶筌により，学習データ中の文を単語レベルにまで分割する．次に，単語ごとに肯定的な文と否定的な文のどちらに出現しやすいかという情報をスコア化する．文を分類する際には，文中の単語のスコアを手がかりにして分類する．なお，本研究では，使用する単語を以下に挙げるものに限っている．

名詞-一般	形容詞-自立
名詞-サ変接続	形容詞-非自立
名詞-副詞可能	連体詞
名詞-ナイ形容詞語幹	副詞-助詞類接続
名詞-形容動詞語幹	副詞-一般
動詞-自立	接頭詞-名詞接続語句
動詞-非自立	

ただ，以上に挙げた品詞においても，品詞によってはノイズとなる単語の方が多い場合もあるので，本研究では，品詞ごとに重みを与えている．品詞の重み付けに関する手法は 3.1.1 節で詳しく述べる．また，使用する学習データはあらかじめ肯定か否定のラベルがふられたものであり，表 1 はその例である．

3.1.1　スコアリング

提案する分類器は，単語ごとに肯定的な文と否定的な文のどちらに出現しやすいかという情報をスコア化して利用する．また，単語そのものにつけられたスコ

表 1: 評価文とラベル付けの例

評価文	ラベル
低価格で持ち運びに適した PC を探しているユーザにお薦めできる	肯定
ハイスペックで非常に使いやすいですが，音が気になります	否定

アに対し，品詞による重みをかけることにより，ノイズとなりやすい品詞に属する語の影響を弱くしている．式 2 は単語にスコア付けをおこなうための式である．

$$Score(w_i) = W_{pos(i)}(P_i^P - P_i^N) \tag{2}$$

式 2 中の P_i^P と P_i^N は tf・idf の概念を基にして算出した，単語 w_i の肯定的な文，否定的な文への出現しやすさを表した数値である．

tf・idf とは，Web ページの検索等によく用いられる手法であり，通常，tf(term frequency) は，対象となる文書においてある単語がどれくらいの頻度で出現するかを表したものである．tf が大きいほどその単語がその文書を良く特徴付けていると言える．また，idf(inverce document frequency) はその単語が出現する文書数が少ないほど，その単語がよく特徴づけていると考えるものである．tf・idf は単に tf と idf を掛け合わせたものである．本手法では，この考え方を利用し，以下の式で P_i^P と P_i^N を求める．

$$P_i^P = \frac{tf_i^P \left(N^P + N^N\right)}{df_i N^P} \tag{3}$$

$$P_i^N = \frac{tf_i^N \left(N^P + N^N\right)}{df_i N^N} \tag{4}$$

本研究では，対象となるクラスに出現する単語 w_i の頻度を tf としている．ここでのクラスとは肯定のクラスと否定のクラスの 2 つである．つまり，式 3 の tf_i^P は肯定的な文の集合に出現する単語 w_i の頻度であり，式 4 の tf_i^N はその逆である．また，ここでの df_i は，両方のクラスにおいて，単語 w_i が出現した文の数としている．N_P，N_N はそれぞれ肯定・否定の文に含まれる単語の総数であり，これらを用いて，肯定・否定間のスコアバランスを調整する．

式 2 の $W_{pos(i)}$ は品詞による重みを表したものである．学習データ中で，同じ品詞に属するそれぞれの単語が，肯定もしくは否定のどちらかのクラスに偏って出現しているかを判定し，偏りのある単語の割合によって，品詞による重みを決定する．ここで，偏りに有意性があるかどうかは，χ^2 検定によって求め，最終的な重みは，式 5 によって求める．

$$W_{pos(i)} = \frac{\text{偏りに有意性がある単語の数}}{\text{その品詞に属する単語の数 (種類)}} \tag{5}$$

χ^2 検定 (chi-square test) は，変数が正規分布に従う乱数かどうかを調べる検定である．本研究では，式 6 に示す χ^2 値が有意水準 α を越えるか否かで判定する．

$$\chi^2 = \sum_{i=1}^{k} \frac{|O_i - E_i| - 0.5}{E_i} \tag{6}$$

ここで，O_i は実際の観測値，E_i は予想される期待値である．k は実測値のクラス数である．本研究における O_i は，対象となる単語がそのクラスに出現した頻度であり，E_i は，肯定のクラスと否定のクラスに出現する全ての語の比率を基に導き出したもので，対象となる単語がそのクラスに出現する頻度を予測したものである．また，k は肯定と否定の 2 クラスである．なお，5 節の実験では，有意水準は一般的に使用される 5%とした．

3.1.2 分類の方法

入力された文を分類する場合には，入力文中にある単語のスコアの総和を文のスコアとし，その正・負によって文を分類する．

3.1.3 その他の処理

基本的な処理はこれまでに述べたとおりであるが，その他の細かい処理として，本研究では，学習・分類の両方において，文脈を考慮した処理などをおこなっている．この節ではこれらの処理について述べる．

逆接の接続詞の処理

逆接の接続詞によって文節がつながれていた場合，接続詞より前の部分のスコアを反転させる．

否定の助動詞の処理

後ろに否定の助動詞を持つ単語は，スコアを反転させる．

名詞化

形容詞や形容動詞の直後に接尾語「さ」が付く語は，名詞としてあつかう．例えば,「鮮やか」と「鮮やかさ」では，前者の場合，単独でも肯定的な評価している語となりえるが，後者は，その後に付く語によって評価が異なってくる．そこで，この 2 つの語の差別化をおこなう為に，このような処理をおこなう．

同義語

学習データ中に同じ読みで同じ品詞の単語があれば，そのスコアを代用する．例えば,「綺麗」と「奇麗」

の場合，漢字が違うだけで，読み方も意味もほとんど同じである．このように同じ読みであれば同じ意味をもつ語である可能性が高いために，このような処理をおこなう．

4 ラベルなしデータを利用した学習データの拡張

これまで，意見情報や評判情報を含む文を肯定的な文と否定的な文にうまく分類するための手法について述べてきたが，評価文分類の処理において，その精度は学習データの内容や量にも大きく依存する．場合によっては，分類手法を改良するより，学習データの量を増やすだけのほうが精度が向上することもある．なぜならば，学習データの量が増えれば，信頼性や網羅性が向上し，精度が向上すると考えられるからである．しかしながら，人手で大量の学習データを作成するには過大なコストがかかる．そこで，ここでは，人手によるコストを抑えつつ，学習データを拡張させる方法について述べる．

4.1 学習データ拡張

本手法では，あらかじめラベル付けされた学習データを基にして，大量のラベル無しデータから自動的に学習データを獲得する．以下に手順を示す．

1. 3 節で述べた分類器をラベル付きデータで学習する．
2. 1 をラベル無しデータに適用し，肯定・否定に分類する．
3. 分類された文のうちスコアの絶対値が高かったものに対し，ラベルを付与し学習データに追加する．
4. 1〜3 を N 回繰り返す．

ここで，最も適切な N を求めるための簡単な実験をおこなったところ，N の値を変化させても，それほど大きな差が出なかった．このため，5.3 節で紹介する実験では，実験の効率を良くするために，N を 1 として処理をおこなう．

4.2 信頼性と網羅性

学習データの量を増やす目的は，信頼性の向上と網羅性の向上の 2 つである．4.1 節で述べた手法で学習データを増やした場合，網羅性の面では向上が見込めるが，新たに獲得した学習データは，元の学習データより，信頼度の面で劣るのは明らかである．元の学習データに新たに獲得した学習データを追加した場合には，データの数が増えることによる信頼度向上という利益とともに，信頼度の低いデータを追加することによる信頼度低下のリスクも負うことになる．

そこで，網羅性のみを向上させるために，追加したデータに含まれる，新しく出現した単語のみのスコアを追加し，他の単語のスコアは元の学習データのみからスコア付けしたものを使用する手法も提案する．また，新しく出現した単語のみのスコアを追加する場合には，元の学習データと追加した学習データで，母数となるデータ数に違いがあり，スコアのバランスが崩れるため，式 7 で求める重みによって調整する．

$$W = \frac{\text{元の学習データ数 (文)}}{\text{新たに追加するデータ数 (文)}} \tag{7}$$

5 節で，全ての単語のスコアを更新する場合と，新しい単語のスコアのみ追加する場合とで，比較実験をおこない，どちらの手法がよいかを検証する．

5 実験と考察

ここでは，本提案手法の有効性について検討するための実験と，4 節で紹介した 2 種類の学習データ拡張手法について検証するための実験をおこなう．

まず，3 つの分類手法による分類実験をおこない，提案手法の有効性を検証する．次に，4 節で紹介した 2 つの学習データ拡張手法を適用した際の分類精度の変化から，2 つの手法の有効性と相違点について検証をおこなう．また，最後に大規模なデータを使って，精度評価と，計算コストの評価のための実験をおこなう．

5.1 使用するデータ

使用するデータは，人手で収集・ラベル付けしたパソコンに関する CNET レビュー記事 2200 文 (肯定：1100 文，否定：1100 文) [1]，大規模なウェブデータから自動構築したコーパス約 50 万文 (肯定：約 22 万文，否定：約 29 万文) [2]，そして，人手で収集したラベル付けされていないパソコンに関するレビュー記事 (価格.com)14763 文[3] の 3 種類である．

大規模なウェブデータから自動構築したコーパスには，意見情報，評価情報を持つ文のみが存在し，自動的に，肯定か否定の評価極性ラベルがふられている．なお，自動構築されたコーパスであるため，人手でラベル付けされたものより，ラベル付けの精度は落ちると考えられるが，10 億という膨大な文書数を利用し，厳しい制限のもとで獲得されたデータであるため，人手のものに近い精度があると考えてよい．

5.2 分類器の精度比較

ここでは，CNET2200 文を用いて，提案手法による分類実験，SVM による分類実験，藤村らの手法による分類実験をおこない，提案手法の有効性を検証する．また，文脈を考慮した処理による影響もみるため

[1] www.sbpnet.jp/vwalker/review/index_c.asp?page=1
[2] http://www.tkl.iis.u-tokyo.ac.jp/ kaji/acp/
[3] http://http://kakaku.com/prdevaluate/newreview.aspx/

に，3 つそれぞれの手法において，文脈の処理がある場合とない場合で精度比較をおこなう．なお，SVM と藤村らの手法における品詞制限などの条件は提案手法と同じものとする．

実験は，10 分割交差検定によりおこなった．10 分割交差検定とは，全体の 9/10 を訓練データ，残りの 1/10 をテストデータとし，それを 10 回繰り返して実験し，その平均を精度とする方法のことをいう．実験結果の F(F-measure)，A(Accuracy) は，以下の計算式 (式 5.2，式 9) を利用した．

$$F = \frac{2 \times P \times R}{P + R} \tag{8}$$

$$A = \frac{\text{正解した文数}}{\text{すべての文の数}} \tag{9}$$

また，式中の P(Precision)，R(Recall) は，以下の計算式 (式 10，式 11) で与える．

$$P = \frac{\text{正解した文数}}{\text{そのクラスに割り当てた文数}} \tag{10}$$

$$R = \frac{\text{正解した文数}}{\text{そのクラスに割り当てるべき文数}} \tag{11}$$

表 2 は提案手法，SVM，藤村らの手法の 3 つの手法の分類実験結果である．

表 2: 3 つの手法の分類精度

手法	文脈	肯定 F	否定 F	A(%)
提案手法	なし	77.81	76.33	77.09
	あり	**82.78**	**81.83**	**82.32**
SVM	なし	76.78	77.03	76.91
	あり	75.70	77.16	76.45
藤村らの手法	なし	76.92	76.44	76.68
	あり	81.23	80.30	80.77

この表より，提案手法と藤村らの手法において，文脈による処理が大きく精度に作用することが分かる．SVM においては，文脈の処理をおこなっても，ほとんど精度は変化せず，むしろ悪影響を与えてしまっているため，スコアリングによる手法では，文脈などの処理をおこなうなど，手を加えることで精度を向上させやすいというメリットがあることが分かった．また，文脈等の処理の有無に関わらず，提案手法が最も良い結果を得られているため，提案手法で用いたスコア計算式が有効であったことが分かる．

5.3 学習データ拡張手法による精度変化

ここでは，4 節で紹介した 2 つの学習データ拡張手法を適用した際の分類精度の変化から，2 つの手法の有効性と相違点について検証をおこなう．

5.3.1 2 つの拡張手法の精度比較

実験は，ラベルなしデータから獲得するデータを徐々に増やしながら，10 分割交差検定をおこなうことにより，それぞれの手法での精度変化をみた．交差検定に用いたデータは CNET データ 2200 文であり，それとは別に，ラベルなしデータとして価格.com データ 14763 文を使用した．図 1 は，全ての単語のスコアを更新する場合と，新しい単語のスコアのみ追加する場合の実験結果をグラフにしたものである．グラフの精度は A(Accuracy) で評価したものであり，グラフ中の破線は，学習データ拡張手法を使用しない場合の精度である．

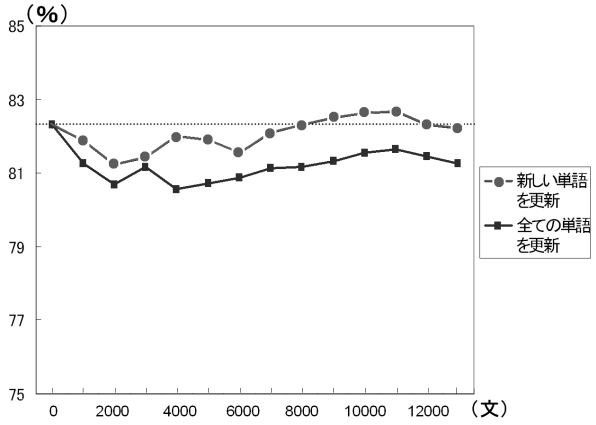

図 1: 2 つの手法の比較

図 1 を見ると，全ての単語のスコアを更新する場合では，獲得するデータを増やしても，元の精度を上回ることはなかった．また，新しい単語のスコアのみを追加する場合では，元の精度を上回ることもあるが，ほとんど変わらない程度であった．2 つの手法を比較すれば，新しい単語のスコアのみ追加した方が良いことは明らかであるが，これら 2 つの手法の有効性に関しては，判断の難しい結果となった．

また，新しい単語のスコアのみを追加する場合の精度変化を見てみると，獲得データが少ないときには，精度が低下するが，その後，徐々に向上している．さらに，獲得データを増やすと，精度は頭打ちとなり，

最終的にやや下落傾向になる．初めに精度が下がる理由は，獲得データが少ないと，新しく出現した単語をスコア付けするのに使用する学習データが少ないということになるので，新しい単語が上手くスコアリングができていないためであると考えられる．獲得数を増やしていくと，新しい単語のスコアリングで使える情報量が増え，精度は徐々に上昇する．その後，精度が頭打ちになることについて，考えられる理由として，2つほど挙げられる．一つは，学習データ拡張手法があまり有効でなく，この精度が能力的な限界であった可能性である．もう一つは，学習データ拡張手法の能力には関係なく，そもそも，元の学習データにおいて網羅性が十分であった為，学習データ拡張手法によって網羅性を高めることが難しかった可能性である．これらの可能性については，5.3.2 節と，5.3.3 節で検証する．精度が最終的に下落傾向になる理由としては，今回の実験で使用したラベルなしデータの数に対して，獲得する文数を増やしすぎたため，極性を持たない文まで分類し学習データに追加してしまったためであると考えられる．

5.3.2　学習データ数の違いよる精度比較

ここでは，前節で述べた 2 つの可能性について検証する．

まずは，元の学習データの網羅性を検証する．これを検証するには，元の学習データの網羅性を変化させて実験をおこなえば良い．ここでは，網羅性を変化させるために，元の学習データの数を変えて実験をおこなう．なお，学習データの網羅性の違いによる変化をみる実験であるため，データ拡張手法は適用しない．

実験は，10 分割交差検定を少し変更した方法でおこなった．学習データの数による精度変化をみるために，通常の 10 分割交差検定で，全体の 9/10 を訓練データ，残りの 1/10 をテストデータとするところを，今回の実験では，1/10 をテストデータとするのは同じだが，他の 9/10 のうち，学習するデータを，0/10, 1/10, …, 8/10, 9/10 と変化させて学習する．例えば，少ないデータで学習させる際には，通常，全体の 9/10 を訓練データとするところを，全体の 3/10 のみで学習させるようにする．これを 10 回繰り返して実験し，その平均を精度とする．

図 2 は，全体の精度 (Accuracy) を折れ線グラフで表したものである．グラフを見れば一目瞭然であるが，学習データが増加すると精度が向上した．また，その伸び具合は対数関数のような伸び方をしており，この後さらに学習データを増やせたとしても，精度は

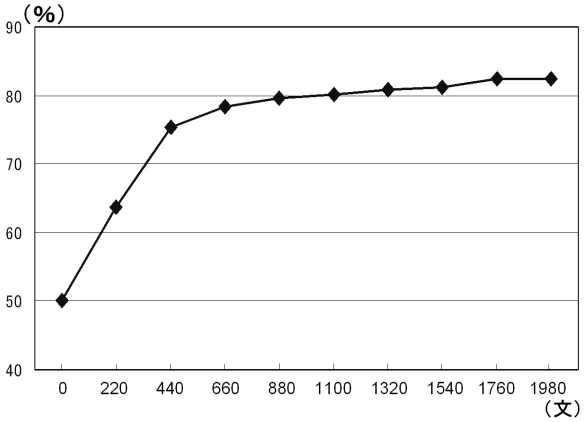

図 2: 学習データによる精度変化

85%前後で収束しそうである．この結果，5.3.1 節の実験における元の学習データの網羅性が，十分高かったと結論づけることができる．

5.3.3　学習データ拡張手法適用時の精度変化

次は学習データ拡張手法の能力をみるための実験である．もし，網羅性の低いデータを使用したとしても，精度が向上しないようであれば，拡張手法そのものに問題があるといえる．

前節の実験は，学習データ拡張手法を適用せずに，学習データの数を変化させて精度の違いを検証したが，今回は，前回実験したそれぞれの学習データ数ごとに，学習データ拡張手法を適用し，精度変化を検証する．実験した学習データ数は，全体の 3/10，5/10，7/10，9/10 の場合である．実験は前節と同様，10 分割交差検定を少し変更した方法でおこなった．交差検定に用いたデータは CNET データ 2200 文であり，それとは別に，ラベルなしデータとして価格.com データ 14763 文を使用した．

学習データ数ごとの，学習データ拡張手法適用結果を図 3 に示す．グラフ中の精度は A(Accuracy) で評価したものである．

グラフ中の破線は，それぞれの学習データ数における学習データ拡張手法を使用しない場合の精度である．

図 3 を見ると，学習データ数が全体の 7/10，9/10 の場合は，獲得する学習データ数を増やしても，元の精度とほとんど変わっていないが，3/10，5/10 の場合は，獲得する学習データをある程度増やせば，元の精度を上回っていることが分かる．このことより，元の学習データの網羅性が低い場合には，提案した学習データ拡張手法が有効であるといえる．

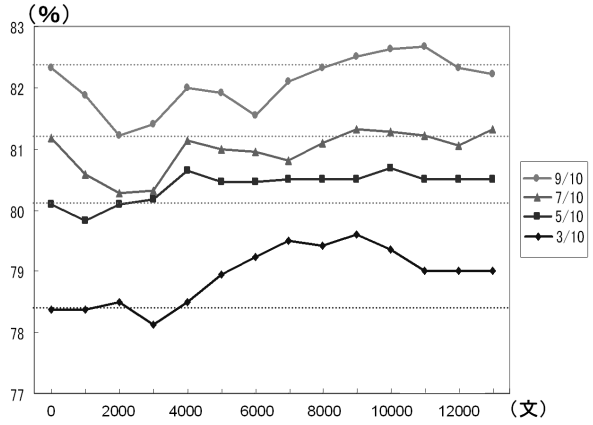

図 3: 学習データごとの精度変化

5.4 大規模データによる精度

ここでは，Web 上の大規模データから自動獲得した約 50 万文のコーパスを利用して，大規模データによる分類精度の検証と，学習分類にかかる時間から，コスト面での検証をおこなう．比較のために，提案手法と SVM で 10 分割交差検定による実験をおこなった．表 3 は提案手法，SVM の分類実験結果である．

表 3: 2 つの手法の分類精度

手法	肯定 F	否定 F	A(%)
提案手法	**80.73**	**85.93**	**83.75**
SVM	72.62	77.84	77.63

表 3 より提案手法が格段に良い精度であったことが分かる．この実験は，使用したデータが異なるだけで，5.2 節の実験と同じ条件でおこなわれたものである．表 2 と表 3 を比べると，両手法ともに全体の精度 Accuracy が約 1%ずつ上昇している．この約 1%の変化は，学習データの増加量からすれば，大した差ではないと考えられる．このことから，CNT データ 2200 文の実験で学習データとして用いた 1980 文は学習データの量としては，十分な量であるといえる．さらに，学習データとして十分であるとするならば，網羅性の面でも十分な量であるといえる．つまり，5.3.3 節の実験では，学習データが全体の 9/10(1980 文) や 7/10(1760 文) の場合には既に飽和状態に近い状態であり，学習データを拡張する必要がなかったため，精度が向上しなかったと考えられる．

また，SVM による分類実験に要した時間は，提案手法による分類実験の 30 倍ほどであった．これにより，計算コストの面でも，提案手法の方が有利であるといえる．

6 分類した知識の利用

ここでは，スコアリングによる手法を用いて分類することによって可能となる利用法の例を紹介する．

- ある製品のレビューに含まれる文を分類器で分類し，肯定的な文，否定的な文の割合から、製品の良し悪しを推測する
- ある製品のレビューに含まれる文を肯定と否定に分類し，その文の中に含まれる単語を提案した手法でスコアリングすることで，単語ごとのスコアにより製品の特徴を把握する

以上の 2 つの応用例の有効性について検証する．今回，検証で使用するデータは，価格.com から収集した 2 つのノートパソコンに関するレビュー記事である．一つ目の製品は VAIO VGN UX50 で，レビューに含まれる文の数は 106 文であった．もう一つは daynabook AX 940LS PAAX940LS で，レビューに含まれる文の数は 56 文であった．それぞれの製品についての，レビューを書いたユーザらによる評価を図 4 に示す．

図 4: ユーザ評価

これら 2 つの製品に関するレビュー記事中の文を今回提案した分類器で肯定的意見と否定的意見に分類し，肯定的な文が占める割合を調べたところ，VAIO VGN UX50 が 55.06%で，daynabook AX 940LS PAAX940LS が 74.55%であった．図 4 中の全体的な満足度は，daynabook AX 940LS PAAX940LS の方が高く，肯定的な文が占める割合はこれと同じ傾向が表れたことになる．評価文書中の肯定的な文と否定的な文の比率が，その文書の評価と高い相関を持つ

表 4: 単語ごとのスコア

製品名	項目	スコアを見た単語	スコア	ユーザ評価
VAIO VGN UX50	デザイン	「デザイン」	**1.433** (+)	**4.0** (+)
	処理速度	「速度」	**0.007** (+)	**3.5** (+)
	グラフィック性能	「グラフィック」	スコアなし	————
	拡張性	「拡張」	**-0.349** (−)	**3.0** (−)
	使いやすさ	「機能」	**-2.132** (−)	**2.8** (−)
	携帯性	「携帯」	**0.542** (+)	**4.0** (+)
	バッテリ	「バッテリ」	**-2.284** (−)	**2.2** (−)
	液晶	「液晶」	**1.535** (+)	**4.0** (+)
	満足度	「満足」	**1.433** (+)	**3.6** (+)
daynabook AX 940LS PAAX940LS	デザイン	「デザイン」	**1.224** (+)	**4.5** (+)
	処理速度	「速度」	-0.186 (−)	3.8 (+)
	グラフィック性能	「グラフィック」	**1.277** (+)	**3.6** (+)
	拡張性	「拡張」	-2.985 (−)	3.9 (+)
	使いやすさ	「機能」	スコアなし	————
	携帯性	「携帯」	**-5.719** (−)	**2.7** (−)
	バッテリ	「バッテリ」	**-0.911** (−)	**2.5** (−)
	液晶	「液晶」	-0.479 (−)	4.1 (+)
	満足度	「満足」	**1.224** (+)	**4.1** (+)

ことは，Pang らの調査 [4] によっても示されているが，我々の提案手法によって機械が分類した文においても，その調査と同じ結果が得られることが分かった．

次に，上で分類した文中の単語に対し，本論文で提案した手法を用いてスコアリングをおこなった．これを用いて，図 4 中のそれぞれの評価項目に関連する単語のスコアごとにその特徴を推測できるかを調べた．

表 4 は項目ごとの単語のスコアとユーザらが 5 段階で評価したスコアの平均との比較である．表 4 において，スコアの後にある ( ) は，単語のスコアが正であった場合に (+)，負であった場合に (−) としており，ユーザ評価の後にある ( ) は，ユーザらが 5 段階で評価したスコアの平均が基準値より良かった場合に (+)，基準値より悪かった場合に (−) で表している．また，単語のスコアの極性がユーザ評価の極性と一致していた場合に太字で表示している．基準値は，ユーザの評価の最高値 4.5 と最小値 2.2 の中間をとった 3.35 としている．

表を見ると，単語のスコアの正・負と，ユーザの評価が一致していたのは全 16 項目中 13 項目であった．この結果から，単語のスコアは製品の特徴をある程度捉えているといえる．また，VAIO VGN UX50 の場合，100%の確率で一致していたのに対し，daynabook AX 940LS PAAX940LS は，63%程度である．これは，daynabook AX 940LS PAAX940LS のデータ数が少なかったことと，文の割合が肯定に偏りすぎていたことが原因として考えられる．

7 おわりに

本論文では，Web 上の掲示板などの文章から収集した意見情報や評判情報を含む文を肯定的意見と否定的意見に分類することで，情報の内容把握をより容易にすることを目的とした研究について紹介した．

本研究では，スコアリングを基に評価文を分類する手法を提案した．実験により，提案手法が SVM や他のスコアリングによる手法よりも高い精度を得られることを実証し，その有効性を確認した．そして，文脈を考慮した処理をおこなうことで精度が格段に良くなるという結果を得た．また，本論文では，学習データを自動的に増やすことで，分類器の精度を向上させることを目指し，元の学習データを使って自動的に学習データを拡張させる手法を提案した．実験の結果，提案手法は元の学習データが少ない場合に有効であるという結果を得た．最後に，分類した評判情報を利用した応用例として，提案した分類器によって分類した文を利用して，製品を評価する方法について提案し，その有効性を示した．

参考文献

[1] 乾孝司, 奥村学 (2006). ”テキストを対象とした評価情報の分析に関する研究動向” 自然言語処理, Vol.13, No.3, 2006.

[2] Pang, B., Lee, L., and Vaithyanathan, S. (2002). ”Thumbs up? Sentiment Classification sing Machine Learning Techniques.” In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), pp. 76-86.

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

[4] Pang, B. and Lee, L. (2005). ”Seeing Stars: Explitting Class Relationships for Sentiment Categorization with Respect to Rating Scals.” In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005).