



評判情報のレベルを考慮した評価文書の分類と評価 情報の信頼性評価への応用

安村, 禎明
坂野, 大作
上原, 邦昭

(Citation)

自然言語処理, 14(3):297-313

(Issue Date)

2007-04

(Resource Type)

journal article

(Version)

Accepted Manuscript

(URL)

<https://hdl.handle.net/20.500.14094/90001008>



評判情報のレベルを考慮した評価文書の分類と評価情報の信頼性評価への応用

安村 禎明[†] 坂野 大作[†] 上原 邦昭[†]

本論文では、Web上の評判情報を有益に活用するために、レビューなどの評価文書をポジティブ(おすすめ)とネガティブ(おすすめしない)という極性値に分類する手法を提案する。本手法では、全体評判情報と部分評判情報という2つのレベルで評判情報を捉える。全体評判情報とは評価文書の対象全般に関わる評価表現のことを指し、部分評判情報とは対象の一属性に関する評価表現のことを指す。全体評判情報の極性値は評価文書の極性値と一致すると考えられるため、まず全体評判情報を用いて評価文書を分類し、全体評判情報がない場合は部分評判情報を用いて分類する。これら2つのレベルの評判情報を考慮することで分類精度の向上が期待できる。さらに、これら2つのレベルの評判情報を用いることで、評判情報の信頼性評価の一手法を提案する。ここでは、評価文書の極性値とその中の部分評判情報の極性値が異なる場合にその部分評判情報は信頼性が高いと評価する。映画のレビューを用いた評価実験の結果、ナイーブベイズを用いた分類手法よりも本手法の方が良い結果が得られた。また、提案した評価指標が評価文書の信頼性評価の一つとなりうることを示唆した。

キーワード: 評判情報, テキスト分類, 信頼性, 情報抽出

Classification of Feedback Documents Considering a Level of Reputations and Reliability Evaluation of Reputations

YOSHIAKI YASUMURA[†], DAISAKU SAKANO[†] and KUNIAKI UEHARA[†]

This paper describes a method for classifying feedback documents into two polarities: positive and negative. In this method, we classify reputations into “Object Level Reputations” and “Attribute Level Reputations”. Object level reputations are the reputations concerning the object of the feedback document. Attribute level reputations are the reputations concerning one attribute of the object. Since we assume that the polarity of the object level reputation corresponds to the polarity of the feedback document, feedback documents are first classified using the object level reputation. The feedback documents that do not contain object level reputations are classified using the attribute level reputations. In addition, this paper proposes a method for reliability evaluation of reputations considering two levels of reputations. In this method, we regard the attribute level reputations that has the opposite polarity of the feedback document as reliable reputations. The experimental results using movie reviews showed that the proposed method could classify feedback documents more correctly than the previous method, and that the proposed measure can be one of the reliability measures for reputations.

KeyWords: *Reputation, Text Classification, Reliability, Information Extraction*

[†] 神戸大学大学院工学研究科情報知能学専攻, Department of Computer and Systems Engineering, Kobe University

1 はじめに

近年、Webが爆発的に普及し、掲示板等のコミュニティにおいて誰もが容易に情報交換をすることが可能になった。このようなコミュニティには様々な人の多様な評判情報(意見)が多く存在している。これらの情報は企業のマーケティングや個人が商品を購入する際の意思決定などに利用されている。

このため、このような製品などに対する評判情報を、Web上に存在するレビューあるいはブログなどから、自動的に収集・解析する技術への期待が高まっている。このため、従来このような評判情報の抽出に関して研究されてきた (Morinaga 2002; 飯田, 小林, 乾, 松本, 立石, 福島 2005; Dave, Lawrence, Pennock 2003; 鍛冶, 喜連川 2005; 矢野, 目良, 相沢 2004; 鈴木, 高村, 奥村 2004)。これらの研究では、製品などに関する評価文書から自然言語処理技術を用いて評判情報を抽出する。また、評判情報を含む評価文書を、ポジティブ(おすすめ)とネガティブ(おすすめしない)という2つの極性値に分類し、その結果をユーザに提示する。提示された情報を基にユーザは様々な意思決定を行う。

評価文書を2つの極性値に分類する手法に関して、これまで多くの研究が行われてきた。(Turney 2002)では、フレーズの極性値に基づく教師なし学習によって評価文書を分類している。(Chaovalit 2005)では、映画のレビューを対象に教師なし学習 (Turney 2002) と教師あり学習を比較している。ここでは、教師あり学習として N-gram を用いている。実験の結果、分類精度は教師あり学習の方が高かったと報告している。教師あり学習を用いたものとして (Dave 他 2003) では、ナイーブベイズを用いて評判情報の分類学習を行っている。

これらの研究では、文書に含まれている単語や評判情報をすべて同等に扱っている。しかし、評価文書には、全体評判情報と部分評判情報という2つのレベルの評判情報が含まれていると考えられる。全体評判情報とは、評価文書の対象全般に関わる評価表現のことを指す。例えば、映画のレビューにおいて「この映画はおもしろい」という評価表現は対象全般に関わる評価表現であり、この表現がある場合はその極性値が評価文書の極値にほぼ一致する。一方、部分評判情報とは、対象の一属性に関わる評価表現のことを指す。例えば、映画のレビューにおいて「映像がきれい」という評価表現は映画の一属性である映像に関する評価表現であり、この表現があったとしてもその極性値が評価文書の極性値と一致するわけではない。したがって、これら2つのレベルを考慮することで評価文書の分類精度の向上が期待できる。

そこで本論文では、評判情報を全体評判情報と部分評判情報という2つのレベルに分け、その極性値を基に評価文書を分類する手法を提案する。本手法では、まず評価文書から全体評判情報を抽出し、その極性値を判定する。この極性値は評価文書の極性値とほぼ一致するため、この極性値を評価文書の極性値とする。評価文書に全体評判情報が含まれない場合は、部分評判情報の極性値の割合から評価文書の極性値を決定する。

さらに、この2つのレベルの評判情報を用いて、評判情報の信頼性を評価するための一手法

を提案する。評判情報は主観的な情報のため、信頼性が低いという問題点がある。このため、その信頼性を評価できれば有益な情報となる。信頼性を評価する手法は多くのことが考えられるが、ここではその1つとして、評価文書の極値と異なる極性値を持つ部分評判情報は信頼性の高い情報と捉えることを提案する。例えば、「すごく面白い映画だった。映像も素晴らしかった」と「はっきりいって最低の映画でした。でも映像だけは良かったです」という評価文書があるとする。前者のように、映画全体をポジティブに評価している人が映像に関してもポジティブに評価することはあまり情報としての価値はない。悪意のある見方をすると宣伝ともとれる。一方、後者は映画全体としてはネガティブな評価であるが、映像に関してはポジティブに評価している。このような評価は客観的でフェアである可能性が高いため、信頼性が高い評価情報であるとする。このような信頼性は、評判情報の2つのレベルを用いることで評価できる。

2 評判情報

評判情報と評価文書を定義し、その表現法について述べる。また、全体評判情報と部分評判情報という評判情報の2つのレベルに関して述べる。

2.1 評判情報と評価文書

Web上ではブログや掲示板あるいはレビュー等で映画の感想やある製品に関する評価が多く存在する。例えば「映像に迫力がなかった」というような文がある。このような評価を含む情報を本研究では評判情報と呼ぶ。また、評判情報を含む文書を評価文書と呼ぶ。レビューで考えれば、1投稿が1つの評価文書に対応する(図1)。

また、「こんにちは、私は原作は見てないので比較はできませんが、この映画はとても面白かったです。」という文章は、評判に関わる部分は「この映画は面白い」ということである。必要な部分だけを抽出すると、評判情報は(対象, 属性, 評価表現)という3つ組の形で表すことができる(飯田他 2005; 立石, 石黒, 福島 2001)。対象とは評価対象の名前や評価対象全体を表す言葉である。属性とは評価対象の一部分を表す言葉であり、評価表現とはその評価対象や属性の評価である。映画を例にした評判情報の表現として、(映画, 映像, 迫力-ない), (ϕ , 演技, すごい), (映画, ϕ , 面白い)などが挙げられる。 ϕ は、実際のデータでは対象や属性が必要ない場合や、文章中で省略されている場合を表す。本研究では、この三つ組みは文単位で抽出する。

評判情報には、大別して全体評判情報と部分評判情報という2つのレベルが存在すると思われる。全体評判情報とは評価対象全般に関する評価表現であり、部分評判情報とは評価対象の一属性に関する評価表現である。例えば、(映画, ϕ , おもしろい)は評価対象全般である映画に関する評価表現であるため、全体評判情報である。一方、(ϕ , 映像, きれい)は映画の一部である映像に関する評価であるため、部分評判情報である。また、(映画, 映像, きれい)のような

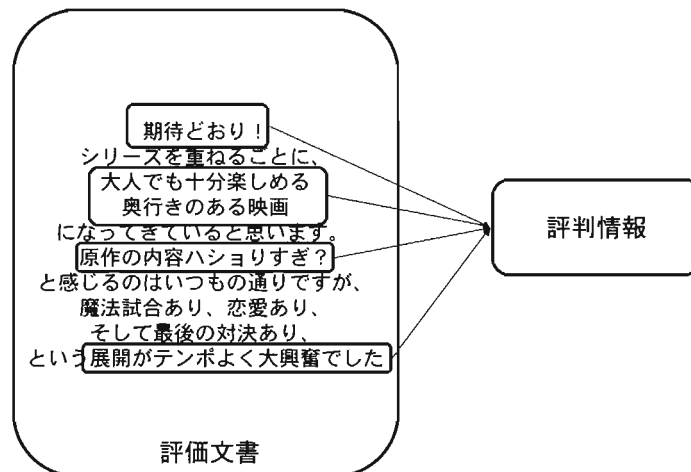


図 1 評価文書と評判情報

すべての属性がある場合でも、映像に関する評判情報であるため、部分評判情報と捉える。全体評判情報は対象全般に関する評価であるため、その極性値は評価文書の極性値と一致すると考えられる。一方、部分評判情報は一属性に関する評価であるため、その極性値は評価文書の極性値と一致するとは限らない。したがって、全体評判情報を重視して評価文書を分類すると、その分類精度の向上が期待できる。

本研究において上記の3つ組の表現を用いることで全体評判情報と部分評判情報を明確に区別することが可能となる。また、本研究では、レビュー内のテキスト情報以外の情報は用いていない。例えば、投稿者の情報や投稿の返信関係、文を超えた範囲からの抽出などは行っていない。これらの情報を考慮することで評価文書の分類精度が上がるのが期待できるが、本論文では評判情報の2つのレベルに焦点を当てているため、これらの情報を考慮しなかった。

3 評価文書の分類

ここでは、評価文書をポジティブとネガティブに分類する手法を説明する。まず、本手法での基本モデルとして用いるナイーブベイズ (NB) モデルについて述べる (Dave 他 2003)。次に、2つのレベルの評判情報を用いた評価文書の分類手法を提案する。

3.1 ナイーブベイズモデル

文書分類では単語の順序は必ずしも必要ではなく、文書中にどのような単語がどのような頻度で出現するかの情報で十分な場合が多い。そこで、単語の順序を無視し、文書を単語の集合として捉える BOW (bag-of-words) モデルが用いられる。BOW モデルでは、1つの文書は形態素解析によって抜き出された単語リスト $d\{w_1, w_2, \dots, w_M\}$ と表現され、単語リスト d は文書と同一視される。 w_m は文書に含まれる単語で、各々は異なる単語とは限らない。この考えに基づき、分類する文書 (実際には単語リスト) を d 、分類するための手がかりとなる学習コーパス (N 個の学習用文書) $D = \{d_1, \dots, d_N\}$ から形態素解析等の処理によって得られる単語リストを、単語集合 W とする。単語集合は $W = \{t_1, t_2, \dots, t_V\}$ と表現する。 t_i は第 i 番目の単語で V は単語の総数を表す。

NB モデルでは BOW モデルに従う。NB では、あるトピック c を持つ文書 d の各単語 w_m の生起を統計的に独立と仮定しているため、独立性の定義から次の式が成り立つ、

$$p(d|c) = P(w_1, \dots, w_M|c) = \prod_{m=1}^M P(w_m|c) \quad (1)$$

これはあるクラスを与えたときに文書 d が生成される確率は、 $w_M \in W$ の生成確率である $P(w_m|c)$ の乗算で算出できることを意味する。

次に単語頻度ベクトル $x = (x_1, \dots, x_V)$ を導入する。 x_i は $t_i \in W$ が文書 d に出現する回数を表す。単語 t_i ごとに整理すると、 $P(w_1|c) \times \dots \times P(w_M|c) = P(t_1|c)^{x_1} \times \dots \times P(t_V|c)^{x_V}$ が成り立つため、 $P(t_i|c) = \theta_i$ とすると、式 (1) は次のようになる。

$$p(d|c) = \prod_{i=1}^V \theta_i^{x_i} \quad (2)$$

これが BOW モデルに基づく、文書の NB モデルである。 $\theta = (\theta_1, \dots, \theta_V)$ は未知のパラメータであるために学習する必要がある。本研究では、NB のパラメータ学習に、事後分布最大化学習 (MAP 学習) を用いる。MAP 学習では、与えられた学習用コーパス D に対して、 θ の事後分布 $p(\theta|D)$ を最大化するパラメータを最適としている。MAP 学習による θ の推定値は次の式で表現できる。

$$\hat{\theta} = \frac{\sum_{n=1}^N x_{n,i} + \alpha - 1}{\sum_{i=1}^V \sum_{n=1}^N x_{n,i} + V(\alpha - 1)} \quad (3)$$

ここで、 $x_{n,i}$ は t_i が D_i 中に出現する頻度ベクトルである。推定パラメータ α は一種の平滑化 (スムージング) パラメータである。 $\hat{\theta}_i$ は W 中の全ての単語が D 中に出現する総回数に対する、 t_i が D 中に出現する数の割合となっている (図 2)

NB モデルを用いて文書をポジティブとネガティブの 2 つのクラス ($c_1 = P, c_2 = N$) に分類する。各クラス毎に学習データ D から式 (3) を利用して $P(d|c_1)$ と $P(d|c_2)$ が得られる。クラスが未知の文書 d^* に対して、クラス事後確率 $P(c_i|d^*)$ を最大化するクラス c_i がベイズ誤り確率最小化

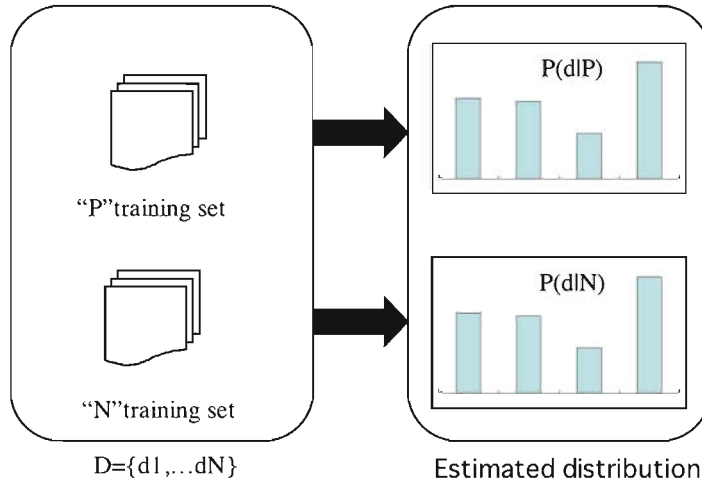


図 2 ナイブベイズモデル

の観点で最適なクラス分類となる。ここで、 $P(c_i|d*) \propto P(c_i)P(d*|c_i)$ なので、 $P(c_i)P(d*|c_i)$ の最大化となる。

3.2 NB 分類器の作成

単語素性の NB 分類器は次の手順で作られる。

- (1) 学習データから評価表現候補単語リストの作成
- (2) NB モデルの作成

学習データ D には、極性値がラベル付けされたレビュー集合を用いる。 D から、評価表現の単語リストを抽出する。このとき、評価表現候補は、形容詞-自立、動詞-自立、名詞-形容動詞語幹、名詞-サ変接続、副詞-一般、副詞-助詞類接続という品詞で絞り込んだ単語集合を抽出する。これは評価表現はこれらの品詞にほぼ限定できるためである。この段階では明らかに評価表現でない名詞を多く含んでいる。そこで、評価対象を特徴づける対象名や属性名 (映画ならば“映画”や“映像”など) と助詞-連体化や助詞-並立助詞、つまり“の”や“と”で繋がる名詞を D から抽出し、評価表現ではない可能性が高いので評価表現候補からは除外する (表 3)。また、明らかに評価表現にならない動詞も候補から外す。これには“する”などの stop-word と呼ばれるものが含まれる。次に、評価表現候補を TF-IDF で得点づけする。TF-IDF は単語を得点づけするアルゴリズムで、ポジティブである評価文書だけに頻出するような単語は TF-IDF 値が大きくなる。逆に、ポジティブにもネガティブにも出現するような評判情報は TF-IDF 値が押さえられ

る．この段階で、ポジティブとネガティブそれぞれに対して特徴的な語が数値として得られる(図1)．

ポジティブ、ネガティブそれぞれに特徴的な語を数値順でソートしたものの中で上位の単語だけを用いる．これは、あまり数値が低いものは特徴的な単語でない可能性があるためである．また、候補数に差があると、分類に偏りが出易くなるので、ポジティブとネガティブ、それぞれの候補数を合わせる．本研究では、ポジティブとネガティブで候補数が少ない方に数に合わせた．このようにして単語集合 W が完成する．さらに単語集合 W の D 中における頻度ベクトル X を作成し、式(3)に基づき $P(d|N)$ と $P(d|P)$ を作成する．

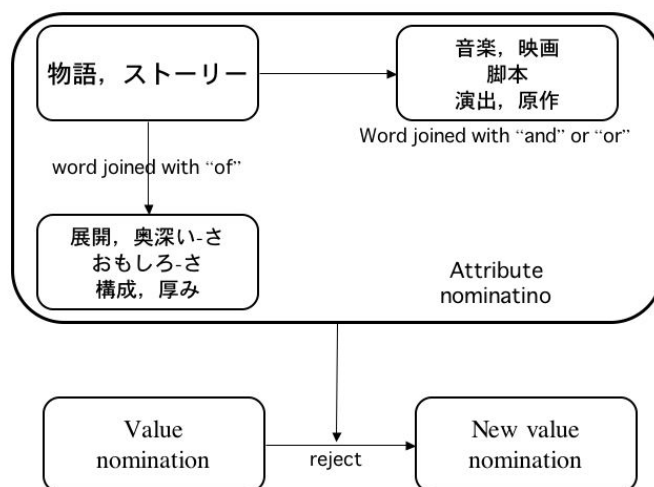


図 3 属性と属性評価表現の抽出

3.3 評価文書分類の提案モデル

ここでは、評価文書を分類するための提案モデルについて述べる．NBモデルによる分類には、分類対象を評価文書にした場合には以下のような問題点がある．問題点の1つは係り受けを扱えないことである．例えば、「車がはやい」と「電池切れがはやい」のように、評価表現だけでなく、対象と評価表現、あるいは属性と評価表現の組でなければ、ポジティブとネガティブに正確に分類できない．このため、係り受けを扱うことで分類精度が高められると期待できる．もう一つの問題点は、全体評判情報と部分評判情報を同等に扱っていることである．単なる多数決ではなく、レベルの違いを利用した分類手法が必要である．例えば、(映画, ϕ , おもしろ

表 1 TF-IDFによるスコアリング

Positive word	tf-idf	Negative word	tf-idf
楽しめる	963	悪い	786
最高	920	良い	634
面白い	873	ない	617
楽しい	845	面白い	377
いい	625	飽きる	371
良い	598	面白い-ない	361
とても	563	つまらない	332
すごい	484	感じる	318
ない	435	もっと	302
すき	434	拾う	300
おもしろい	433	ちょっと	298
よい	387	わかる-ない	298
演技	345	好き	258
監督	345	がっかり	251
笑う	344	無い	248
感じる	326	どう	243
楽しむ	323	正直	242
ちょっと	310	変	236
もう一度	300	残念	223

ろい)が一回, (ϕ , 映像, 荒い)が2回出現するような評価文書は一般的にはポジティブに分類する. なぜならば, 全体評判情報の極性値は評価文書の極性値とほぼ一致するため, これを重視すると, この評価文書の極性値はポジティブであると予想されるからである. しかし, 評判情報の単純な多数決ではこの評価文書はネガティブに分類される. したがって, 全体評判情報を重視すれば, 評価文書の分類精度は向上すると期待できる.

評価文書を分類するための提案モデルには次の2つの新しい点がある.

- 全体評判情報と部分評判情報に分けて文書分類すること
- 係り受けの関係を扱うこと

提案モデルでは, 次の2つの分類器を作成する. 1つは全体評判情報の分類器, もう1つはNB分類器である. 最初に全体評判情報分類器で分類を試みる. この分類器では全体評判情報を抽出し, その極性値を求める. 全体評判情報の極性値は評価文書の極性値とほぼ一致するため, 評価文書を全体評判情報の極性値に基づき分類する. 全体評判情報を含まない評価文書はこの分類器では分類不可能であるため, このような評価文書は, 単語素性のNB分類器を用いて分類する. この手順により本手法では優先的に全体評判情報を用いて評価文書を分類する. 実際に全体評判情報として使われる素性は評判情報の3つ組の種類のうち(対象, ϕ , 評価表現)となる.

3.4 2つのレベルを考慮した評価文書の分類手法

全体評判情報の分類器の作成手順を以下に示す.

- (1) 学習データ D から対象候補単語リストの作成
対象候補単語はその対象の特徴的な言葉に限定し，人手で設定する．例えば，映画ならば“映画”，“作品”を用いる．
- (2) D から評価表現候補単語リストの作成
単語素性に基づく NB 分類器の評価候補単語リストの作成と同様に行う．
- (3) D から対象候補と評価表現候補の組み合わせとのマッチングによる全体評判情報候補を作成
 D から係り受けの関係にある 2 文節をすべて抽出し，対象候補と評価表現候補の組み合わせとマッチングしていく．このようにして抜き出されたものを同じ組み合わせであるもの毎に集めて単語集合 W と単語集合の頻度ベクトル X を得る．(図 4)
- (4) NB モデルの作成
 W と X と式 (3) から $P(d|N)$ と $P(d|P)$ を作成する．

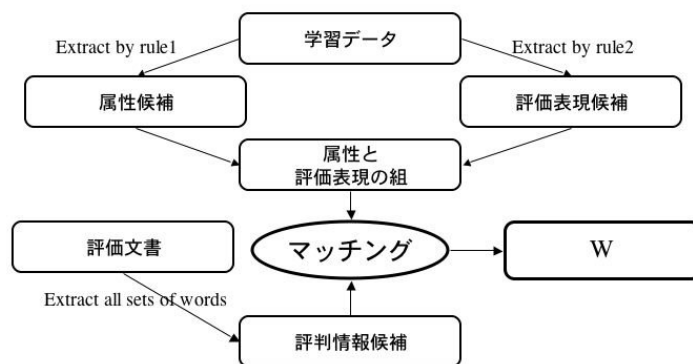


図 4 属性と評価表現の抽出

4 評価情報の信頼性評価

ここでは，評価情報の信頼性を評価する一手法について述べる．評判情報は主観的な意見であるために，その客観性は乏しく信頼性は低いという問題点がある．このような信頼性の低い情報の中から比較的信頼性の高い情報を抽出できれば有益な情報となる．

人間は主に2つの信頼性評価方法を用いていると考えられる。1つは、サイト名などの情報発信者や組織名などの情報を用いる方法である。このような情報を元に情報の信頼性を評価する研究(Mui 2002; 竹原, 中島, 角谷, 田中 2004)があるが、サイトの信頼性が低いからといって全ての評判情報の信頼性が低いわけではない。また、投稿者情報を用いることは匿名性の高さのために困難であることが多い。もう1つの信頼性評価手法は、複数の情報の整合性から評価するものである。これを利用した研究は単純に多数決をとることで客観性を与えるということ(立石他 2001)しかなされていない。

本論文では、信頼性評価の一要素として評価文書と評判情報の極性値に基づく手法を提案する。ここでは、評価文書の極性値とその中の部分評判情報の極性値が異なる場合にその部分評判情報は信頼性が高いと評価する。例えば、「すごく面白い映画だった。映像も素晴らしかった」と「はっきりいって最低の映画でした。でも映像だけは良かったです」という評価文書について考える。前者のように、ポジティブな極性値を持つ評価文書において、ポジティブな部分評判情報はあまり情報としての価値はない。悪意のある見方をすれば、前者の評判情報は映画の宣伝とも捉えられる。しかし、後者は映画そのものはネガティブと捉えているが、映像に関してはポジティブに評価しているため、客観的でフェアな評判情報と考えられる。このように、評価文書の極性値とは逆の評価を持つ部分評判情報は、他のものよりもフェアであると考えられる。この理由は以下の通りである。

- 対象全般に対する評価と属性に対する評価が同じになることが一般的であるが、あえて異なる極性値を持つ評価情報を書き込むことは情報としての価値が高い。
- 対象に関してポジティブな面とネガティブな面の両方が評価できているため、客観性が高い。
- 宣伝や熱狂的なファン、アンチファンの投稿は信頼性が低いですが、このような情報を排除できる。

本論文では、このような情報をフェアな評判情報と呼ぶ。

このフェアな評判情報を抽出するためには、評価文書と評判情報の極性値も調べることが必要である。つまり、評価文書を分類するタスクと、評価文書から評判情報を抜き出した後、各評判情報を分類するタスクの2つのタスクが必要である。(図5)

この2つのタスクの結果、部分評判情報は以下の4種類に分類される。

- 評価文書としてはポジティブ、部分評判情報としてはポジティブなもの(Pp)
- 評価文書としてはポジティブ、部分評判情報としてはネガティブなもの(Pn)
- 評価文書としてはネガティブ、部分評判情報としてはポジティブなもの(Np)
- 評価文書としてはネガティブ、部分評判情報としてはネガティブなもの(Nn)

フェアな評判情報はPnとNpということになる。本論文では、このように評価文書を分類し、その中の評判情報を分類することでフェアな評判情報とそれ以外の評判情報を区別する。

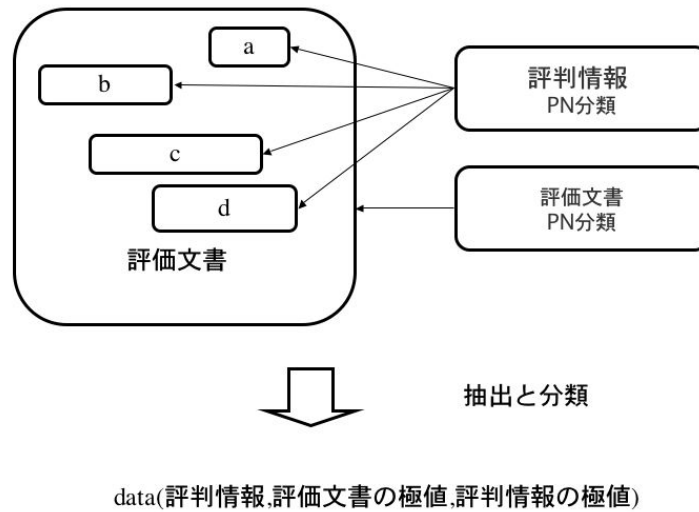


図 5 信頼性評価手法の概要

このようにしてフェアな評判情報を抽出する。

4.1 フェアな評判情報の抽出

本節では、フェアな評判情報の抽出法について述べる。フェアな評判情報を抽出するためには、評価文書の分類と、それに含まれている評判情報の抽出およびその分類が必要である。評価文書の分類に関しては、前章で提案した手法を用いる。以下では評判情報の抽出とその分類について述べる(飯田他 2005)。

このタスクにおける評判情報とは部分評判情報を指すため、(ϕ , 属性, 評価表現)を抽出し、分類するタスクである。このタスクのために、まず、評判情報辞書を作成する。基本的な考え方は辞書にマッチする評判情報候補は評判情報であるというものである。この手法を用いる理由は、既存の研究では様々な条件付けで評判情報候補を絞ることはできても、それが実際に評判情報であるかという分類は難しいとされているためである。

辞書の作成には、まず学習データ D から属性候補と評価表現候補を抽出することから始める。属性候補に関しては、初期値として属性であると考えられる単語を10程度与える。学習データ D 中でそれらと助詞-連体化や助詞-並立助詞、つまり「の」や「と」で繋がる名詞を抽出する。このように抜き出された名詞は、対象の属性である可能性が高いため初期値にこれを加え属性候補とする。評価表現候補は、形容詞-自立、動詞-自立、名詞-形容動詞語幹、名詞-サ変接続、副詞-一般、副詞-助詞類接続という品詞で絞り込めるため、これらの品詞を候補とす

表 2 評判情報の分類

	Pp	Pn	Np	Nn
カテゴリ 1	14	0	86	100
カテゴリ 2	145	22	0	123

る．このように抽出された属性候補と評価集合の全組み合わせに対して手動でポジティブとネガティブをラベル付けして，正事例として辞書に加える．評価表現でないと判断した組み合わせは負事例として学習していく．このように作成された評判情報辞書を用いることで，分類対象の評価文書から評判情報を抜き出す．

4.2 フェアな評判情報の利用

本手法によって得られた部分評判情報を評価対象毎，カテゴリ毎，ラベル (Pp や Nn) 毎にカウントすることによって，表 2 のような分類結果が得られる．カテゴリとは属性のグループであり，映画のカテゴリでは映像・音楽・ストーリーなどである．このカテゴリと属性を一对多で対応させることで，カテゴリ毎，ラベル毎に集計する．

この表をユーザに提示することで評判情報が評価できる．例えば，この表でのカテゴリ 1，カテゴリ 2 とともに単なるポジティブとネガティブの多数決をとると，それぞれ 100 対 100 と 145 対 145 で同じとなる．しかし，フェアな評判情報である Pn と Np を考慮することで，カテゴリ 1 はポジティブが優勢であり，カテゴリ 2 はネガティブが優勢であると判断できる．

5 評価実験

評価文書分類において NB モデルと提案手法の比較実験を行った．また，抽出されたフェアな評判情報の有用性について評価する．

5.1 実験設定

実験に用いたデータはポータルサイト “Yahoo Japan” の “Yahoo Movie” から収集した．収集したデータは，最近公開されたメジャーな 10 タイトルにおいてそれぞれ最新の 1000 レビューを収集したものであり，合計 10000 レビューである．これらのタイトルを選択したのは，レビュー数が十分であることと，実際に本手法を用いる際には比較的新しい情報を対象とすることが多いと考えたためである．“Yahoo Movie” のレビューには，投稿者によって点数が 5 段階で付与されている．本実験では得点が 5 点あるいは 4 点のレビューをポジティブとし，2 点あるいは 1 点のものをネガティブとした．3 点または “得点なし” は中立とした．図 3 にデータの詳細な内訳を示す．

表 3 実験データ

データ	ポジティブ	ネガティブ	合計
タイトル 1	647	242	889
タイトル 2	426	356	782
タイトル 3	592	207	799
タイトル 4	750	149	899
タイトル 5	812	122	934
タイトル 6	790	131	921
タイトル 7	547	261	808
タイトル 8	426	396	822
タイトル 9	654	283	937
タイトル 10	542	298	840
合計	6186	2445	8631

10回交差検定でNBモデルと提案手法の結果を比較した。データの中から、1000レビュー、つまり1タイトル毎に評価データとし、残りのデータである9000レビューを訓練データとした。比較の指標には精度と再現率を用いた。また、評判情報に関する語は自立語に絞られるため、両手法共に素性には自立語のみを利用した。また、否定語である“ない”に関しては、“自立語-ない”の形で扱っている。全体評判情報の最初の対象候補単語リストとしては、映画、作品、ムービーを用いた。部分評判情報の最初の候補単語リストは、映像、CG、画面、音楽、ミュージック、曲、演出、ステージング、脚色、配役、キャスティング、キャスト、物語、ストーリー、話を用いた。

5.2 全体評判情報による分類評価

ここでは、全体評判情報が実際に評価文書分類に有用であるかの評価のために、全体評判情報を人手で抽出し、全体評判情報と評価文書の評価がどの程度一致するかを調べた。ここでの人手による全体評判情報の抽出は、「良かった」、「最悪」などの一単語で明確に評価が分かるものに限定し、「ちょっと...」、「心に残る」などの分かりにくい表現は抽出に用いなかった。

500のレビューを人手で評価した結果、全体評判情報は178のレビューに含まれており、そのうち162のレビューでは全体評判情報と評価文書の極性値が一致した。また、一致しなかった16のレビューに関しても、評価文書の評価値はすべて3点となっており、中立の評価となった。したがって、全体評判情報によって逆の極性値を取るものはなかった。全体評判情報が含まれている文書は全体の1/3以上であり、全体評判情報が一般的な情報であることがわかった。以上のことから、全体評判情報が評価文書分類に有用であることが示唆された。更に、提案手法において全体評判情報がNBよりも高い精度で抽出できれば評価文書の分類精度が向上することが予測される。

表 4 評価文書の分類結果

	P 精度	P 再現率	N 精度	N 再現率	全体精度	全体再現率
NB モデル	0.832	0.7943	0.538	0.578	0.741	0.733
提案モデル	0.848	0.799	0.570	0.639	0.771	0.760

5.3 評価文書分類の実験結果

評価文書分類の実験結果を表 4 に示す。値は 10 回交差検定の平均値である。P 精度、P 再現率とは、それぞれポジティブな評価文書に対する精度、再現率を表しており、N 精度、N 再現率、全体精度、全体再現率はそれぞれネガティブな評価文書と全評価文書に対する精度と再現率を表す。人間がこの分類を行った場合、9 割強の精度で分類できるが、完全には分類できないと思われる。この理由は内容に関して述べているだけで評価につながる表現がないレビューやポジティブとネガティブの両方の評価が書いてあるが、結局全体としてどちらに評価したのかわからない場合が挙げられる。

実験結果としては、提案モデルが NB モデルをすべての精度、再現率において上回った。これは全体評判情報を用いて分類したためと考えられる。本手法の方が正確に分類した例として「先日、見てきました。とても面白かったです。ですが、レンのお腹……。あれはないかと……。ちょっと失笑してしまいました。もう少し役作りして欲しかったです。」というレビューがある。このレビューでは、「とても面白かった」という全体評判情報によってポジティブな投稿であると本手法では判定しているが、NB モデルでは、それ以外の単語も考慮しているため、この影響でネガティブに分類された。

しかしながら、全体の精度と再現率に関しては NB モデルと提案手法の t 検定による有意差はなかった。この原因は、明らかに全体情報でない候補に対して強い特徴づけをしていることが主として挙げられる。本手法による全体評判情報の抽出精度は 82% であり、全体評判情報をさらに高い精度で抽出する必要がある。全体評判情報の抽出精度を上げるには辞書の拡充があげられるが、人手によるコストが大きくなる。

また、ネガティブな評価文書に関する性能の低さは、ポジティブと比べて、学習データが少ないことが原因として考えられる。これは、学習データ数を揃えることで解決できそうであるが、ネガティブな評価文書数は表 3 からわかるように少なく、揃えることが容易ではない。また、日本人の特徴である“否定的な事ははっきり言わない”ということを考えると、数を揃えるだけでは対処できない場合もある。例えば「ちょっと……」のような表現がある。

分類誤りは、Yahoo の得点とは逆の内容を書いている、ということを除けば以下のような例がある。

- 全体評判情報の候補リストに誤っているものが含まれている場合
例えば、(映画, ϕ , 聞く)(映画, ϕ , 感じる)などがネガティブな候補として上位となった。

これらの全体評判情報は極性値を決定するものではないが、分布の偏りによってはこのように全体評判情報の候補となる。このリストを基に分類器を作成するため、分類精度を下げることになった。

- 他人のレビューを引用して否定している場合
“「最高の映画」なんて書いている人がいるけど”のように逆の見地をとる投稿者の評判情報部分を引用していることがある。この問題に対処するためには、引用部分を見分ける必要がある。

5.4 フェアな評判情報の評価

ここでは、フェアな評判情報を評価する。映画の評判を多数決を用いて評価した場合とフェアな評判情報のみを用いて多数決を用いて評価した場合を比較し、どちらが世間的な評価に近いかを確認する。世間的な評価を完全に把握することは困難であるが、ここでは著者の一名が多くのレビューを読むことで世間的な評価を判断した。

実際のデータに対して本手法を適用し、評判情報を分類した。評価データとしてどのように世間的に評価されているかがよく知られているため、“NANA”と“オペラ座の怪人”の2つの映画のレビューを用いた。表5、表6にその分類結果を示す。表中の数値はデータ中に含まれていたそれぞれのカテゴリの評判情報の数である。

NANAの分類結果を見てみると、音楽カテゴリにおいて P_n に対して N_p が多く、ポジティブである可能性が高いと考えられる。オペラ座の怪人の音楽カテゴリにおいても同様のことが言える。実際にこの両作品に関しては音楽的な評価が高かったと考えられるため、世間的な評価と一致している。しかし、これらの結果は単純な多数決でも同様の結果が得られる。

NANAのストーリーに注目すると、単純な多数決をとった場合、224対114となりポジティブの方が多い。しかし、フェアな評判情報である P_n と N_p を比較すると61対72となりネガティブの方が多くなる。NANAのストーリーは世間的には原作とのギャップから評価が低かったことを考えると、ネガティブとする方が妥当である。これはフェアな評判情報のみを用いた場合と一致している。このようにフェアな評判情報を用いることで、世間的な評価を抽出できる可能性を示唆した。このように評判情報の分類結果を用いることで、評判情報の多様な解析が可能となる。

6 おわりに

本論文では、評判情報の2つのレベルを考慮した評価文書の分類手法を提案した。全体評判情報を用いて評価文書を分類し、その後に部分評判情報を用いて分類することによって、分類精度の向上を試みた。映画のレビューを用いた実験の結果、ナイーブベイズによる手法よりも

表 5 NANA の評判情報の分類結果

	Pp	Pn	Np	Nn
映像	9	0	0	0
音楽	98	2	24	2
キャスト	143	15	25	34
ストーリー	163	72	61	42

表 6 オペラ座の怪人の評判情報の分類結果

オペラ座の怪人	Pp	Pn	Np	Nn
映像	65	1	7	0
音楽	172	7	25	10
キャスト	71	8	11	7
ストーリー	147	3	18	2

分類精度が向上することを確認した。

また，評価文書の極性値と評判情報の極性値を利用することで，信頼性の高い情報を抽出するための一手法を提案した．評価文書の極性値と其中的評判情報の極値が異なる場合，その評判情報をフェアな評判情報であるとし，信頼性の高い情報とした．実験により，フェアな評判情報が評判情報を評価する際に 1つの指標となる可能性を示した．今後の課題としては，フェアな評判情報および，評判情報の分類結果から読み取れる情報の利用法があげられる．

参考文献

- P. Chaovalit L. Zhou (2005). "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches ." In *Proceedings of the 38th Hawaii International Conference on System Sciences*.
- K. Dave S. Lawrence D. M. Pennock (2003). "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews ." In *Proceedings of WWW*, pp. 519-528.
- 飯田龍, 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一 (2005). "意見抽出を目的とした機械学習による属性-評価値の同定 ." 自然言語処理研究会, 情報処理学会, vol.165, pp. 21-28.
- 鍛冶伸裕, 喜連川優 (2005). "依存構造を考慮した評価文書の分類 ." 自然言語処理研究会, 情報処理学会, vol.170, pp. 15-20.
- S. Morinaga K. Yamanishi K. Tateishi T. Fukushima (2002). "Mining Product Reputations on the Web ." In *Proceedings of SIGKDD 02*.

- L. Mui A. Halberstadt M. Mohtashemi (2002). "A Computational Model of Trust and Reputaion ." In *Proceedings of the 35th Hawaii Internatioanl Conference on System Sciences*.
- 鈴木泰裕, 高村大也, 奥村学 (2004). "Weblogを対象とした評価表現抽出 ." 人工知能学会, SIGSW&ONT-A401-2.
- 竹原幹人, 中島伸介, 角谷和俊, 田中克己 (2004). "Web 情報検索の為に Blog 情報に基づくトラスト値の算出方法 ." 日本データベース学会 Letters, Vol.3, No1, pp. 101-104.
- 立石健二, 石黒義英, 福島俊一 (2001). "インターネットからの評判情報検索 ." 自然言語処理研究会, 情報処理学会, vol.144, pp. 75-82.
- Peter D. Turney (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417-424.
- 矢野宏実, 目良和也, 相沢輝昭 (2004). "趣向を考慮した評判情報検索手法 ." 自然言語処理研究会, 情報処理学会, vol.164, pp. 165-170.

略歴

安村 禎明: 1993 年大阪大学基礎工学部卒業。1998 年同大学院基礎工学研究科博士後期課程修了。同年東京工業大学大学院助手。2004 年神戸大学工学部助教授。現在、同大学大学院工学研究科准教授。博士 (工学)。機械学習, テキストマイニング, エージェントに関する研究に従事。人工知能学会, 電子情報通信学会, 情報処理学会, IEEE 各会員。

坂野 大作: 2004 年岡山県立大学情報工学部卒業。2006 年神戸大学大学院自然科学研究科情報知能工学専攻修了。現在, NTT コミュニケーションズ株式会社に勤務。在学中, テキストマイニングの研究に従事。

上原 邦昭: 1978 年大阪大学基礎工学部情報工学科卒業。1983 年同大学院博士後期課程単位取得退学。大阪大学産業科学研究所助手, 講師, 神戸大学工学部情報知能工学科助教授, 同都市安全研究センター教授, 同大学大学院自然科学研究科教授を経て, 同大学院工学研究科教授。工学博士。人工知能, 特に機械学習, マルチメディア処理の研究に従事。情報処理学会, 人工知能学会, 電子情報通信学会, 計量国語学会, 日本ソフトウェア科学会, AAAI 各会員。