E-058

評判情報の素性による信頼性の推定
Estimation of reliability by features of reputation

福田 雅志† 佐藤 和† 太原 育夫‡
Masashi Fukuda Kazu Sato Ikuo Tahara

1 はじめに

近年，Webが爆発的に普及し，ブログや掲示板などで誰もが容易に情報を発信することが可能になった．このように発信された情報には様々な人の多様な評判情報が存在している．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多くの研究がなされてきた[1, 2, 3, 4, 5, 6, 7]．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方法や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つに分類し，その結果をユーザに提示する．

これらの研究は，より多くの情報を用いることで客観的な判定を可能にしている．しかし，Web上の情報は玉石混淆であり，評判情報においても様々な意見が存在する．例えば，映画のレビューにおいて「この映画はおもしろい」とただ一言そう書かれている意見と，特徴的な表現やそう思った理由などと共に書かれている場合では，同じポジティブな意見であっても，意思決定をする際のユーザの信頼度は異なると考えられる．そのようなユーザの信頼度を推定できれば評判文書の分類精度の向上が期待できる．

そこで本研究では，ある評判情報がどれほど信頼されそうかを推定して数値で表現することを目的とする．評判情報に関係すると思われる単語数やTF-IDF値などの素性を用いて信頼度を決定する．その結果を人手で与えられた信頼度と比較することで評価を行う．

2 評判情報の信頼性評価

まず本稿で提案する評判情報とその信頼性を評価する既存研究と本手法について述べる．

†東京理科大学大学院理工学研究科情報科学専攻
‡東京理科大学理工学部

2.1 評判情報

Web上ではブログや掲示板，レビュー等の，ある製品に関する感想や評価が数多く存在する．例えば，映画に関する文書で「この映画のシナリオは素晴らしい」というような文がある．このような評価を含む情報を評判情報と呼び，評判情報を含む文書を評価文書と呼ぶ．映画のレビューで考えれば，1つのレビューが1つの評価文書に対応する．

2.2 信頼性評価

人間は，普段から様々な観点で様々な情報の信頼性を評価している．特にWeb上の情報は豊富で玉石混淆であり，評判情報においても様々な意見が存在し，その全ての信頼性を評価することは困難である．ここでは，そのようなWeb上の情報に対して，信頼性を評価する場合にどのような既存研究が行われているかを説明する．

一言にWeb上の情報に対する信頼性と言っても，様々な種類のものがある．まず，サイト名などの情報の発信者や組織名などの信頼性を評価する研究がある[8, 9]．これらは評価する対象のページを言及している他のページを利用して信頼性を決定する手法である．知名度のある企業や人物などに対して，Web上でどのような評価を得る際に有効であると考えられる．しかし，発信元に匿名性がある場合の情報についての評価を得ることはできない．

つぎに，様々な手法を用いて，多くの評判文書から抽出したものを評判情報と同定している研究がある[10, 11, 12]．これらの研究の目的は，特定の商品に対する主な評判情報はポジティブかネガティブであるかを決定することやその要約である．本研究の目的である評判文書の信頼性を決定するものではないが，主な評判情報を多く含んでいる文書は有効な文書であると考えることができる．また，評判情報をさらに詳細に全体評判情報と部分評判情報に分割し，そのポジティブとネガティブの値が異なる場合に信頼性があると考える研究もある[13]．

本研究では，評判文書の信頼性を様々な素性を用いて数値化することを目的とする．上記の研究との相違点は，必ずしも発信者の情報を必要としないため匿名性のある情報にも適用が可能であること，評判情報のみを用いるわけではないこと，様々な素性の中から有効な素性を分析することなどである．

3 信頼性評価に用いる素性

ここでは，信頼性評価に用いる素性とその使用法について説明する．Web 上の情報は話し言葉も多く，また記号などの様々な文字が使用されている．そのため，複雑な解析では誤りが生じ，その後の処理に影響すると考えられる．そこで本研究では，そのような誤りを少なくするために，最も基本的な形態素解析の結果から作成したパターンで素性を検出している．形態素解析器には『茶筌』[14] を用いた．以下に各素性を示す．

- 単語
  形態素解析により分割された数を素性として用いる．長く言及されたことは信頼性があるという観測に基づいている．

- 文
  句点や「!」「?」などの記号や「()」で囲まれた顔文字などの表現，また HTML タグでの改行などによって示される数を素性として用いる．

- TF-IDF 値
  全ての評判情報のテキストを用いて TF-IDF 値を計算し，各テキストの含まれる単語の TF-IDF 値の平均を素性として用いる．

- 名詞
  形態素解析の結果，名詞と判定された数を素性として用いる．

- 形容詞
  形態素解析の結果，形容詞と判定された数を素性として用いる．「面白い」「馬鹿馬鹿しい」などの評判表現が含まれる．

- 形容動詞
  形態素解析の結果，形容動詞語幹と判定された数を素性として用いる．先程と同じく，「綺麗」「重厚」などの評判表現が含まれる．

- 助詞
  形態素解析の結果，助詞と判定された数を素性として用いる．

- 助動詞
  形態素解析の結果，助動詞と判定された数を素性として用いる．

- 記号
  形態素解析の結果，記号の数を素性として用いる．顔文字や「☆」や「♪」，顔文字などの感情表現が含まれる．

- 未知語
  形態素解析の結果，未知語と判定された数を素性として用いる．

- 感動詞
  形態素解析の結果，感動詞と判定された数を素性として用いる．大袈裟な感情表現などが含まれる．

- 接続詞
  形態素解析の結果，接続詞と判定された数を素性として用いる．大袈裟な感情表現を用いた評判情報が含まれる．

- 数値
  数値的な表現の数を素性として用いる．評価するときに点数を付けることや，動員数など客観的な数字を用いた評判情報が含まれる．

評判文書 t において，これらの素性 i_t の値 $x(i_t)$ を用いて，信頼性のスコア $S(t)$ を付与する．

$$S(t) = \sum_i \alpha(i_t) x(i_t) \tag{1}$$

ここで，α_i は，それぞれの素性に対する重みである．

これらの素性は評判文書の信頼度に関係のありそうなものを全て用いており，素性同士の関係は考慮されていない．そのため，後で素性同士の関係を調べる必要がある．

4 信頼度の推定実験

4.1 実験環境

本研究では"YAHOO!JAPAN"の"YAHOO!MOVIE"に投稿された映画のレビューを収集して行った．収集した映画数は全部で約 8000 あり，レビューは総計で約 20 万件あった．その収集したレビューの中から，その数が多い映画 10 作品について実験を行った．その 10 作品におけるそれぞれのレビュー数の平均は約 2000 件である．

各レビューには，役立ち度という「レビューを読んで参考になったと思ったユーザの人数」が記されている．この値は人気のある映画であれば多くなり，異なる映画の場合，同じ人数でもその重みは違う．そこで同タイトル内で最大の値を用いて正規化した．この正規化された役立ち度を以下，正解信頼度と呼ぶ．これに対し，本手法で素性により推定された信頼度を推定信頼度と呼ぶ．

また，役立ち度は 1 つのレビューに対して同じ数の人が観て判断したわけではない．役立ち度が 0 人の場合，多くの人がそのレビューを読んで信頼度が低いと判断したのか，信頼度は高いのだが，まだ誰もそのレビューを読んでいないのかの区別はできない．そこで，今回の実験ではある程度のユーザが読んだと考えられる役立ち度が 10 人以上のレビューのみを対象として実験を行った．

4.2 素性の分析と選択

本節では重回帰分析を行い，式 (1) のような正解信頼度と素性の重回帰式を作成する．重回帰式を決定する手法は幾つかあるが，本研究では以下の手順で決定している．

表 1: 各素性と正解信頼度の相関係数の値

	正解信頼度
単語	0.33
文	0.89
TF-IDF	0.60
名詞	0.87
形容詞	0.63
形容動詞	**0.91**
助詞	0.79
助動詞	0.26
記号	0.72
未知語	0.23
感動詞	0.59
接続詞	0.90
数値	0.02

1. 正解信頼度への影響が高い素性を選ぶ．それぞれの素性において正解信頼度との単相関係数を調べ，正解信頼度に与える影響の大きさを調べる．

2. 素性間で高い相関関係が見られる場合はどちらか一方を落とす．一般に素性間で相関係数が 0.9 以上ある場合は，正解信頼度への影響が小さい方を落とす．

3. 偏回帰係数がほぼ 0 となるような素性は、役にたたないので落とす．

始めに，各素性と正解信頼度の相関係数を調べる．それぞれのレビュー t における要素 i の値 $x(i)$ と正解信頼度 $R(i)$ の相関係数は以下の式によって導かれる．

$$\frac{\sum_t (i_t - \bar{i})(R(t) - \bar{R})}{\sqrt{\sum_t (i_t - \bar{i})^2}\sqrt{\sum_t (R(t) - \bar{R})^2}} \qquad (2)$$

その結果を表 1 に示す．表 1 より正解信頼度と最も相関関係があったのは「形容動詞」であり，素性として採用する．

次に，各素性同士の相関係数を調べ，「形容動詞」と相関関係が高い素性を落とす．各素性同士の相関係数の表を示す (表 2).「文」「名詞」「助詞」「接続詞」などが 0.9 以上の相関係数であるため，これらは素性として採用しない．

最後に，残った素性で重回帰分析を行い，偏回帰係数を調べる．残った素性の偏回帰係数の値を表 3 に示す．表 3 により，「単語」「未知語」「感動詞」を回帰式の素性から落とす．

残った素性で改めて重回帰分析を行い，重回帰式を作成した (式 3).

$$\begin{aligned} \text{推定推定値} &= 0.1380 * TFIDF + 0.6944 * \text{形容動詞} \\ &+ 0.1011 * \text{助動詞} + 0.1444 * \text{記号 1} \\ &- 0.2743 * \text{数値} + 10.4491 \end{aligned} \qquad (3)$$

表 3: 残った素性の偏回帰係数の値

	偏回帰係数
単語	0.00
TF-IDF	0.13
形容動詞	0.67
助動詞	0.09
記号	0.14
未知語	0.02
感動詞	0.03
数値	-0.26

4.3 信頼度の推定評価

本節では，作成した重回帰式を用いて推定信頼度を算出する．算出された推定信頼度と人手で付与された正解信頼度を相対誤差と相関係数により比較する．

評価として用いたタイトルは，先程重回帰分析を行った 10 タイトル以外の 6 タイトルである．それらの推定信頼度を算出し，正解信頼度と比較した．正解信頼度と推定信頼度の相対誤差と相関係数の結果を表 4 に示す．

表 4: 正解信頼度と推定信頼度の相対誤差

	相対誤差	相関係数
映画 A	0.16	0.33
映画 B	0.13	0.23
映画 C	0.16	-0.14
映画 D	0.14	0.40
映画 E	0.12	0.55
映画 F	0.12	0.02
平均	0.138	0.232

4.4 考察

作成された重回帰式により得られた推定信頼度と人手で付与された正解信頼度との相対誤差は，平均して約 0.14 となり映画による違いはそれ程見られなかった．この結果は，本手法により取捨選択された素性は，映画のタイトルに関わらず有効であることを示している．そのため，この値をさらに用いて計算するような処理には向いているといえる．今回の実験では役立ち度が 10 人以上のレビューのみを解析しているので，逆に役立たないような文書を解析に加えることで，相関係数の精度の向上が望める．

相関係数で比較してみると映画により値は大きく変化した．この結果は，各レビューにおいて推定信頼度が上回ったり下回ったりしていることを示しており，本手法は信頼度順に評判文書を並び替えるなどの処理にはまだ不向きであることを示している．しかし，現在用いている素性との相関が低く，標準偏回帰係数が高い素性を発見することで全体的な精度の向上が望める．

表 2: 各素性同士の相関係数の値

	単語	文	T-I	名詞	形容	形動	助詞	助動	記号	未知	感動	接続	数値
単語	1												
文	0.43	1											
TF-IDF	0.26	0.60	1										
名詞	0.36	0.88	0.66	1									
形容詞	0.37	0.70	0.46	0.69	1								
形容動詞	0.34	**0.91**	0.60	**0.93**	0.66	1							
助詞	0.34	0.81	0.64	**0.95**	0.65	0.84	1						
助動詞	0.24	0.34	0.31	0.37	0.28	0.28	0.40	1					
記号	0.49	0.72	0.48	0.66	0.54	0.70	0.58	0.23	1				
未知語	0.15	0.23	0.28	0.30	0.20	0.23	0.29	0.38	0.21	1			
感動詞	0.19	0.56	0.45	0.58	0.39	0.58	0.54	0.16	0.45	0.15	1		
接続詞	0.39	0.93	0.58	**0.92**	0.63	0.95	0.82	0.36	0.73	0.24	0.58	1	
数値	0.18	0.16	0.33	0.24	0.25	0.09	0.32	0.60	0.01	0.31	-0.04	0.09	1

5 まとめ

本研究では，評判文書の素性を組み合わせて用い，信頼度を推定する手法を提案した．多くの素性の相対係数を調べ有効な素性を選び，重回帰分析を行い推定の式を作成した．そして，人手で与えられた信頼度と比較することによって式を評価した．その結果，映画に関わらず相対誤差は約 0.14 という結果になり，今回の手法で取捨選択した素性は映画に関わらず有効であることを示した．今後の課題としては，有効な素性を発見し誤差をより少なくすることや，この信頼度を利用した手法の提案などが挙げられる．

参考文献

[1] 関根聡: テキストからの情報抽出, 情報処理学会誌, Vol.40, No.4, pp.370-373, 1999.

[2] Dave,K., Lawrence,S., and Pennock,D.M., "Mining the peanut gallery Opinion extraction and semantic classification of product reviews." In Proceedings of WWW, pp.519–528.

[3] 鈴木泰裕, 高村大也, 奥村学: Weblog を対象とした評価表現抽出, 人工知能学会セマンティックウェブとオントロジー研究会 (SW-ONT-A401-02), 2004.

[4] 鍛冶伸裕, 喜連川優: 文構造を考慮した評価文書分類のための確率モデル, DEWS2006, 5A-i4, 2006.

[5] 杉木健二, 松原茂樹: クエリの主観性に頑健な商品検索システム, FIT 2007, E-056, pp.271–274, 2007.

[6] 岡野原大輔, 辻井潤一: レビューに対する評価指標の自動付与, 自然言語処理, Vol.14, No.3, pp.273–295, 2007.

[7] 藤村滋, 豊田正史, 喜連川優: 文の構造を考慮した評判抽出手法, 電子情報通信学会 第 16 回データ工学ワークショップ (DEWS2005), インタラクティブ発表, 6C-i8. 2005.

[8] Mui,L., Halberstadt,A., and Mohtashemi,M.: "A Computational Model of Trust and Reputation", In Proceedings of the 35th Hawaii International Conference on System Science.

[9] 竹原幹人, 中島伸介, 住谷和俊, 田中克己,: Web 情報検索の為の 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, No.1, pp.101–104.

[10] 立石健二, 福島俊一, 小林のぞみ, 高橋哲朗, 藤田篤, 乾健太郎, 松本裕治: 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, 情報処理学会研究報告, NL-163, pp.1–8, 2004.

[11] 飯田龍, 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: 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, 情報処理学会研究報告, NL-165, pp.21–28, 2005.

[12] 小林のぞみ, 乾健太郎, 松本裕治: 意見情報抽出のための評価対象・評価視点間の関係同定, 言語処理学会第 12 回年次大会論文集, pp.65–68, 2006.

[13] 安村禎明, 坂野大作, 上原邦昭: 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, 自然言語処理, Vol.14, No.3, pp.297-313, 2007.

[14] 松本裕治, 北内啓, 山下達雄, 平野義隆, 松田寛, 高岡一馬, 浅原正幸: 形態素解析システム『茶筌』 version2.3.3 使用説明書.