評価情報における嗜好性と評価表現の関連性に関する考察

百瀬　和徳†　　山口　実靖†　　淺谷　耕一†‡

工学院大学†　　工学院大学大学院‡

1. はじめに

ブログや電子掲示板などを通じて誰もが情報を手軽に交換する事ができるようになり，ネットワーク上には膨大な量の情報が蓄積されることとなった．発信された情報の中には，各個人の対象に対する主観的な評価情報が多数存在している．そのため爆発的に増大した情報の中から主観性のある評価情報を抽出し，分析・活用する評価情報分析に注目が集まっている．

この評価情報分析は評価表現とその極性(肯定/否定)を用いて行われるが，評価情報で用いられる評価表現はドメインや嗜好性により異なる．そのため，評価表現抽出や評価表現辞書の作成は容易ではなく，この手続きの困難さを低減させることは重要な課題の一つと考えられる[1]．

本稿では，異なる嗜好性間における評価表現の類似度の算出，および類似度を用いて評価表現辞書の拡張を行うことの有用性の検証を行う．実験として，映画のレビューを用いた評価表現抽出および類似度算出実験を行った．また，求めた類似関係を考慮し類似嗜好間で評価表現辞書を代用することの有効性の検証を行った．

2. 類似度算出

2. 1　算出手法

本研究では，以下の方法により類似度を求める．まず，Web 上に存在する作品に対する評価値のついたレビュー(HTML 文書)をクロールし，形態素解析を行い単語に分割する．

次に，単語ごとに「各評価値レビューでの出現頻度」を求め，極性を式(1)により求める．これにより評価表現辞書を作成する．例えば，高い評価値が与えられた肯定的なレビューに数多く登場する語は肯定語となる．

最後に各ドメインで抽出した評価表現辞書間の類似度を式(2)，(3)，(4)より算出する．

2. 2　極性判定

単語 w の極性を以下の式(1)により定める．Pi は単語 w が評価値 i のレビューに登場する確率である．Ai は，レビューが作品に対して与えた評価値を正規化した値であり，最大値を 1，最小値を-1 とした．本研究では極性判定のための値を肯定度と呼び，高い評価の特徴語を肯定語，低い評価の特徴語を否定語とする．求まる肯定度が 0 より大きければ肯定語，0 未満であれば否定語と判定する．

$$Positive(w) = \frac{\sum_i^{all} (Ai \times Pi(w))}{\sum_i^{all} Pi(w)} \quad (1)$$

2. 3　類似度

異なるドメイン(嗜好)間で抽出した評価表現辞書同士で比較を行い，以下の式(2)，(3)により類似度を定める．異なるドメイン A と B において，各ドメインにおける単語 w の肯定度をそれぞれ $P_A(w)$，$P_B(w)$，各ドメインの全レビュー内における単語 w の登場確率をそれぞれ $E_A(w)$，$E_B(w)$とし，ドメイン AB 間における単語 w の近さ(距離)を $2-|P_A(w)-P_B(w)|$ により求める．単語 w が A と B の両方に存在する場合，式(2)により類似度を算出する．どちらか片方にしか存在しない場合，式(3)により類似度は 0 となる．両ドメインの類似度が高ければ，単語の両ドメインでの肯定度は近い値となり，$Similarity(w)$は高い値をとる．

そして，式(4)により異なるドメイン AB 間の類似度 $DomainsSimilarity(A,B)$ を算出する．類似度の値が大きい程，類似性が高くなる．

$$\begin{cases} Similarity(w) = \left(2-|P_A(w)-P_B(w)|\right) \times \left\{\dfrac{E_A(w)+E_B(w)}{2}\right\} & (2) \\ Similarity(w) = 0 & (3) \end{cases}$$

$$DomainsSimilarity(A,B) = \sum_w^{all} Similarity(w) \quad (4)$$

3　実験

Yahoo!映画 [3]におけるレビュー投稿者が映画のイメージ(嗜好)として選んだイメージワード 10 語の各上位 20 作品内にあるレビュー約 27 万件を用いて類似度算出実験を行った．各イメージワード間における類似度算出結果を図 1 に示す．

次に，算出したイメージワード間の類似度の結果より，イメージワード「悲しい」に対し最高の類似度を得た「切ない」と最低の類似度となった「パニック」を用いて，評価表現辞書の代用実験を行った．まず，イメージワード「悲

Preference in evaluation information and consideration concer ning relation of evaluation expression
† Kazunori MOMOSE , Saneyasu YAMAGUCHI , Koichi ASATANI
Kogakuin University
‡Koichi ASATANI
Graduate School of Kogakuin University

しい」のレビューより得られた評価表現辞書を用いて，イメージワード「悲しい」の映画作品の肯定レビューの検索を行った．

次に，イメージワード「切ない」より得られた評価表現辞書を用いてイメージワード「悲しい」の肯定レビューの検索を行った．

最後に，「パニック」より得られた評価表現辞書を用いてイメージワード「悲しい」の肯定レビューの検索を行った．各方法で得られた肯定レビュー(評価値 4 以上)の確率を表 1 に示す．

図 1 より，イメージワード「恐怖」において類似度が高かったのは「不気味」や「絶望的」であった．同様に「切ない」と「悲しい」，「笑える」と「楽しい」などのイメージワード間でも類似度が高かった．逆に意味合いが遠い「笑える」や「楽しい」，「コミカル」などのイメージワードと「恐怖」や「不気味」，「悲しい」などのイメージワード間では類似度が低くなることが確認できた．また，「パニック」はどのイメージワードに対しても類似性が著しく低いことが分かった．このことから，これら意味合いが近いイメージワード間では高い類似性が得られることがわかった．逆に意味合いが遠いイメージワード間では類似性が低いことが確認できた．しかし，「悲しい」に対して「絶望的」や「不気味」というような意味合いが近いとは言えないイメージワード間でも高い類似性が得られた．このことから，意味合いが近いとは言えない嗜好性間でも，必ずしも類似性が低いわけではないことがわかった．

表 1 より，「悲しい」に対しの最高の類似度を得た「切ない」が「悲しい」で作成した評価表現辞書と近い精度で判定が行えることがわかった．逆に，最低の類似度であった「パニック」で判定を行った場合，その精度は類似度と同様に低くなることがわかった．よって，嗜好間の類似度を考慮して評価表現辞書の拡張を行うことは有用であると思われる．

4 おわりに

本稿では，評価表現辞書の拡張に嗜好性間の類似性を考慮することの有用性を検討するために，レビュー投稿者が映画のイメージとして選んだイメージワードを用いて異なる嗜好間における類似度算出実験を行った．また，類似度算出結果により，異なるイメージワード間の評価表現辞書を代用する実験を行った．これらの実験より映画という 1 つのドメイン内においても，イメージワード間の類似度を考えることで評価表現の極性や意味が変わることが確認でき，類似性を考慮することで評価表現辞書の精度向上が見込めることがわかった．これにより，嗜好間の類似度を評価表現辞書の拡張に用いることが有用であると考えられる．

今後は映画と小説というような異なるドメイン間における類似度の調査を行い，評価表現辞書の拡張手法の検討を行っていく．

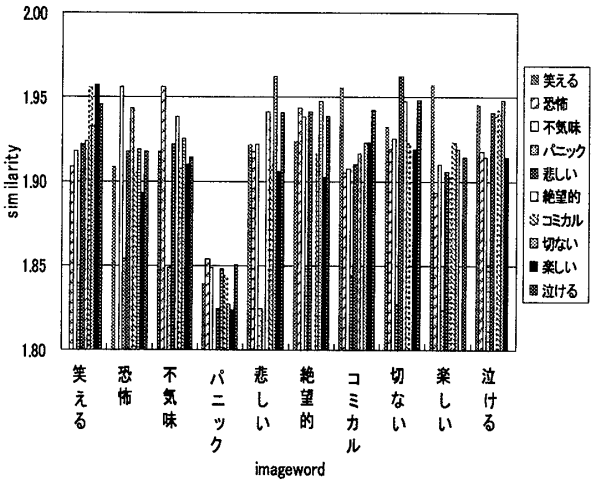

図 1. イメージワード間における類似度

表 1. 類似度を考慮した評価表現辞書代用実験結果

	パニック	切ない	悲しい
悲しい	45.0%	87.5%	92.5%

参考文献

[1]乾孝司, 奥村学, "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, pp. 201-241, 2006.

[2]有安香子, 妹尾宏, 鹿喰善明, "コメントの類似性に基づく視聴者クラスタリング手法の提案", DBWeb2007.

[3]http://movies.yahoo.co.jp/