良質なレビューの特性とそれを考慮した評判情報の抽出

小倉　達矢[*1]　宮野　晋平[*2]　永井　慎介[*2]　山口　実靖[*1]　淺谷　耕一[*1*2]

[*1]工学院大学　[*2]工学院大学大学院

1. はじめに

近年，SNS（Social Networking Service）やブログの利用者の増加に伴い，個人が手軽に情報を発信できるようになった．個人の書き込みなどにはある対象に対しての意見などの評判情報が多く含まれており，これらの情報がネットワーク上に大量に蓄積されるようになった．このようなWeb上の文書から有益な評判情報を抽出できると考えられるが，Web上の情報には品質の保証がなく，抽出情報の高品質化が重要な課題の一個となっている．

本研究では，レビューサイトに存在する意見や感想などの評判情報に付加されているレビュー対する評価に着目し，大衆に支持されているレビューを良質なレビューと仮定した．本稿では良質なレビューの特性の分析と，レビューの品質を考慮した評判情報の抽出を行う．

2. Webマイニングと評判情報抽出

Web マイニングとは Web を対象としたマイニングの手法である．Web 上のデータやテキストなどを解析し，データ間の相関関係やパターンを元に有益な知識や情報を抽出する処理のことである．Web マイニングの研究で代表的なものの一つに評判情報抽出がある．評判情報の抽出とは，レビューやブログから対象に対する意見などを抽出する研究である．Web 上の文書を解析し，意見文の抽出や，各対象物が得ている評価情報を取得する．評価情報の取得では，各文書が肯定的意見であるか否定的意見であるかを判別し，文書の分類などを行う[1-2]．しかし，Web 上のレビュー文には品質の保証がなく，既存の研究ではレビューの文章の品質が考慮されていないという問題点があり[3]，文書の信頼度の取得やこれを考慮した検索手法が提案されている[4]．

3. 良質なレビュー(評判情報)の判別

レビュー文の品質は，以下の 2 手法で決定する．両者を支持率と参考度と呼ぶ．Amazon[5]などのレビューには，「そのレビューを良いと判断した人数」の情報が付加されている．レビューの支持率は式(1)を用いて定義する[4]．

この支持率が閾値以上であるレビューを良質なレビューとする．

$$S = \left(\frac{G}{A}\right) \cdots (1)$$

S=支持率
G=そのレビューを「良い」と判断した人数
A=レビューを参照した人数

本研究では支持率0.7以上のレビューを良質なレビューとし，支持率0.3以下のレビューを低品質なレビューとした．またレビューを参照した人数(上記のA)が5未満のレビューは除外した．

参考度は (参考になった人数)−(参考にならなかった人数) と定める．

4. レビュー文書品質の基本調査

Amazon のレビューを用い，各作品のレビュー郡の「レビューが作品に与えたスコア(1〜5 点)」と「レビューが一般閲覧者から得た評価(支持率と参考度)」を調査した．2 作品の両値の分布を図 1〜4 に示す．図内には全プロットの 1 次近似の式を記した．作品 A では「作品に高いスコアを与えたレビュー」が支持される傾向がある．よって，作品 A は「レビューを評価した一般閲覧者」から支持されていると考えられる．近似直線の傾きは正の値となる．この傾きは「一般閲覧者が作品に対して間接的に与えた評価」と考えることができ，評価が高いほど傾きは大きくなる．以下，これをレビュー傾斜と呼ぶ．作品 B は逆の傾向となり，支持されていないと考えられる．

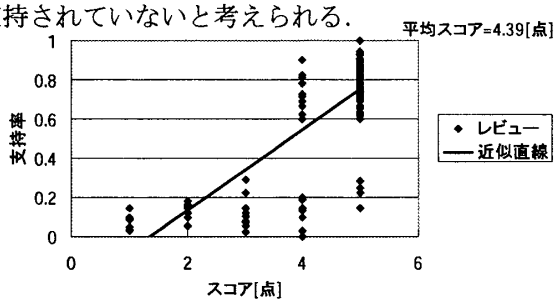

図 1　作品 A のスコアと支持率の分布

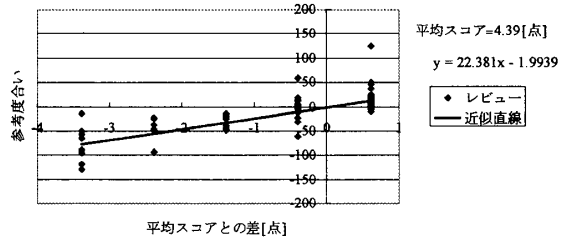

図 2　作品 A のスコアとレビュー傾斜の分布

Opinion Information Extraction Based on Review's Quality
Tatsuya OGURA　Sinpei MIYANO　Shinsuke NAGAI
Saneyasu YAMAGUCHI　Koichi ASATANI
[*1] Kogakuin university.
[*2]Graduate School of Kogakuin University.

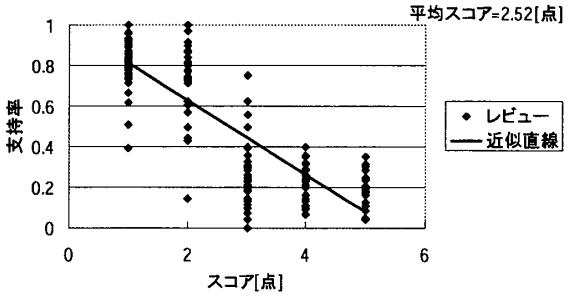

図3　作品Bのスコアと支持率の分布

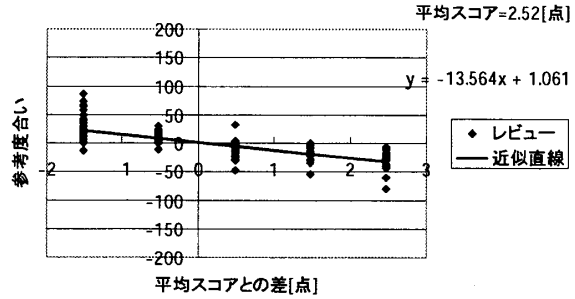

図4　作品Bのスコアとレビュー傾斜の分布

次に，Amazonの DVD レビューサイトの 200 作品(レビュー総数約 15000 件)を対象とし，各レビューが作品に与えたスコアと，各レビューが得た支持率の関係を図5に示す．同図より，高いスコアのレビュー数が多く，高いスコアのレビューが高確率で支持されていることが分かる．低スコアレビューの支持率の分布は高スコアレビューと比べ偏りが少ない．スコアの平均は 3.88 点であり，支持率の平均は 0.48 である．

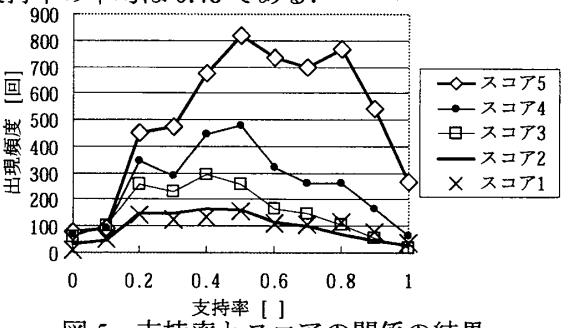

図5　支持率とスコアの関係の結果

5.　レビュー品質を考慮した作品の評価

前章の DVD200 作品の中から 20 作品（レビュー総数約 2000 件）の情報を用い，「全レビューを対象とし作品の評価値を決定する従来の手法」と，「良質なレビューのみを対象とした提案手法」，「低品質名レビューのみを対象とした提案手法」，「レビュー傾斜により作品の評価を定める提案手法」の 4 手法により作品の定量評価を行った．それぞれの手法によるランキング結果を表1に示す．レビュー傾斜手法の評価値は傾きの値を正規化したものであり，それ以外の手法ではそれぞれの手法が対象としているレビューが作品に与えたスコアの平均である．

表1　スコア抽出結果

従来手法

作品	評価値
猟奇的な彼女	4.66
かもめ食堂	4.61
ファインディングニモ	4.55
ロードオブザリング	4.40
いま、会いに行きます	4.34
ALWAYS 三丁目の夕日	4.29
チャーリーとチョコレート工場	4.26
時をかける少女	4.23
硫黄島からの手紙	4.11
ラストサムライ	4.00

良質レビュー

作品	評価値
ロードオブザリング	4.90
猟奇的な彼女	4.86
いま、会いに行きます	4.81
硫黄島からの手紙	4.81
かもめ食堂	4.77
ファインディングニモ	4.76
ALWAYS 三丁目の夕日	4.70
チャーリーとチョコレート工場	4.67
ラストサムライ	4.54
時をかける少女	3.96

低品質レビュー

作品	評価値
踊る大捜査線 THE MOVIE2	4.58
ALWAYS 三丁目の夕日	4.37
ファインディングニモ	4.29
猟奇的な彼女	4.22
千と千尋の神隠し	4.22
時をかける少女	4.16
かもめ食堂	4.13
模倣犯	4.13
ターミナル	4.09
マトリックス リローデッド	4.03

レビュー傾斜

作品	評価値
ロードオブザリング	4.90
チャーリーとチョコレート工場	3.73
硫黄島からの手紙	3.52
時をかける少女	3.26
いま、会いに行きます	3.12
猟奇的な彼女	3.10
かもめ食堂	3.09
ファインディングニモ	3.03
ラストサムライ	3.00
ALWAYS 三丁目の夕日	2.82

表2　主観評価の結果

ランキング	ALL	0.7以上	0.3以下	傾斜
平均評価	4.73	5.40	4.67	6.33

また，提案手法の有効性を確認するために，主観評価実験を行った．レンタルビデオ店の店員 15 人に上記で作成した各ランキングの評価を依頼した．主観評価結果を表 2 に示す(各手法を順に All, 0.7 以上, 0.3 以下，傾き，と記す)．主観評価は 1～10 の点数を付けることにより行い，10 が最も優れた評価である．表 2 より，レビュー品質を考慮することで，全レビューを対象とした従来の手法と比べ，良質な作品の評価(ランキング)が得られることが分かり，提案手法の有効性が確認された．

6.　おわりに

本稿では，Web 上の情報の品質を考慮した評判情報の抽出手法を提案した．提案手法をレビューサイトに対して適用し，主観評価実験を行った結果，より良質な評判情報の抽出が可能であることが確認された．

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,”人工知能学会第18回全国大会,2004

[2]立石健二，石黒義英，福島俊一.”インターネットからの評判情報検索”, 情報処理学会自然言語処理研究会(NL-144-11)

[3]山澤美由紀,吉村宏樹,増市博,”Amazon レビュー文の有用性判別実験,” 情報処理学会研究報告 NL2006-173-(3), pp.15-20, May. 2006.

[4]梅村和宏,鈴木優,川越恭二,“レビューサイトのための批評文の信頼度と支持率を用いた検索手法,”DBWeb2007,パネルディスカッション

[5]http://www.amazon.co.jp/