E-040

映画のブログからの意見情報抽出に基づく関心の分析

Analysis of Concerns from Opinions in Movie Weblogs

甲元 香奈† 関 洋平† 青野 雅樹†
Kana Koumoto Yohei Seki Masaki Aono

1. はじめに

近年，膨大な情報の中から特定の情報に着目し，抽出する技術が注目されている．その中でも，意見に着目することで，多くの人々の意見から，関心の傾向や特徴を抽出することが期待されている．

このような，意見調査を行う手段の一つとしては，アンケートにより多くの人々から意見を収集する方法がある．しかし，多くの意見を人手で収集することは，労力や時間がかかるため，意見を収集し自動的に分類や分析を行う方法が必要である．ブログ記事は，個人の意見や感想を自由に書き発信できるため，率直な多くの意見を容易に収集できる．しかし，ブログ記事は，記述方法が統一されていないため，関心の傾向や特徴を明らかにする項目の設定が必要である．

そこで，本研究では，映画に関するブログ記事に対する複数の側面からの意見の分析を通じて，関心の特徴を表す項目を設定し，多くの人々の関心の傾向を可視化するための方法を提案する．また，その可視化に必要な意見情報を抽出するシステムを実現する．

2. 関連研究

意見抽出・分類に関連した先行研究としては，車などに関する Web 上のレビュー記事中の意見情報から，対象物・属性・評価表現を抽出して可視化を行う手法[1]がある．この手法では，属性ごとの評価を表すレーダーチャートを作成する．

本研究では，映画に関するブログ記事を扱う点が異なる．ブログ記事では，一般にレビュー記事と比べて記述の揺れが大きいと言われているため，関心の特徴を表す項目ごとの分析は難しい．また，映画について書かれたブログ記事を対象とした場合に，鑑賞前と鑑賞後の評価の違いは，映画を見に行くかどうかの判断において重要である．本研究では，この違いの分析に関しても検討している．

3. 映画ブログにおける関心の傾向の可視化

3.1 関心の特徴を表す分類項目の設定

映画の関心の傾向を可視化するためには，映画において関心が集まる特徴を捉える項目が必要となる．本研究では，日本の映画賞である「キネマ旬報」，「ブルーリボン賞」，「毎日映画コンクール」，「日本アカデミー賞」の各賞を参考に，“監督”・“出演者”・“映像”・“音楽”・“ストーリー”の 5 つの項目を設定した．

また，各項目に対して分類のためのキーワードを設定した．キーワードは 4 つの映画についてブログ記事の分析を行い，汎用的に使用できる名詞・動詞を選択した．設定したキーワードの例と設定数を表 1 に示す．

表 1 各項目に分類するためのキーワードの例

項目	数	例
監督	3	監督，演出，メガホン
出演者	34	出演者，主演，演技
映像	31	映像，描写，風景
音楽	20	主題歌，歌，サントラ
ストーリー	56	脚本，物語，展開

このキーワードを用いて，分析に使用した以外の 3 つの映画についてのブログ記事中の記述を各項目に分類した結果，精度はほとんど悪化しなかったことから，これらのキーワード群は個々の映画作品に依存しないと考えられる．

3.2 可視化のための軸の設定

多くの人々の関心の傾向を把握するためには，まず各項目に対して肯定する意見が多いのか，否定する意見が多いのかという評価が重要になる．また，各項目に対して関心を持っている人々の割合を提示することも重要である．

そこで，本研究では可視化するために 2 つの軸を用意した．1 つ目は「評価」という軸である．各項目についての肯定または否定の傾向を明らかにするために，肯定表現をプラス，否定表現をマイナスとし，その差を表す．2 つ目は「言及度」という軸である．言及度は式(1)を用いて計算する[1]．

$$\text{言及度} = \frac{\text{項目}a\text{の出現した記事数}}{\text{収集した全記事数}} * 100 \qquad (1)$$

3.3 可視化結果の確認

映画の関心の傾向が正しく可視化できることを確認するため，7 つの映画に対して，感想の書かれた記事を各 30 記事用意した．また可視化した結果から，関心の傾向を適切に把握できるか，以下の 2 つの手順で確かめた．

(1)正解となるデータとして，被験者 1 名が記事中の 3,789 文について，3.1 節で設定した 5 つの項目の情報と，その各項目に対する肯定・否定の情報を非排他的に付与した．

(2)次に(1)で付与したデータから関心の傾向を可視化し，映画の傾向を適切に表しているか確認した．

図 1，図 2 に可視化の例を示す．図 1 の映画では，「ストーリー」と「映像」の項目に対する意見情報が多く存在し関心が集まっているが，「ストーリー」の評価は悪いことが分かる．この結果は“ドラマを詰め込みすぎ”，“つまらない”といった意見が影響している．図 2 の映画では，「ストーリー」と「出演者」の項目について意見情報が多く存在するが，「ストーリー」よりも「出演者」の評価が高いことが分かる．この結果は，主演俳優への肯定評価が多いことが影響している．

† 豊橋技術科学大学 工学研究科 情報工学専攻
kana@kde.ics.tut.ac.jp, seki@ics.tut.ac.jp, aono@ics.tut.ac.jp

[1] すなわち，多くの記事に同じ項目についての評価が現れていれば，多くの人々がその項目に関心を持っていると考える．

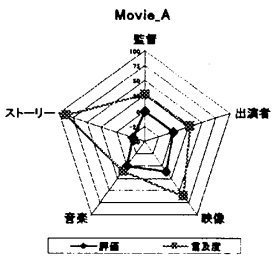

図 1　可視化の例 1

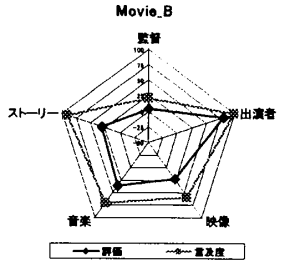

図 2　可視化の例 2

また，鑑賞前と，鑑賞後の評価についても可視化を試みた．今回収集した記事 210 記事中 52 記事(3,789 文中 55 文)に鑑賞前の期待する評価が書かれた文があった．以下に，鑑賞前の評価が記述されていた文の例を示す．

- 予告編からして面白そう！
- 正直あまり期待してなかったんだけど，
  思いのほか面白かった．

この鑑賞前と鑑賞後の評価の，肯定と否定の割合を可視化することで鑑賞により評価が変化する様子を示す．図 3 に正解と設定したデータによる評価が変化した例を示す．

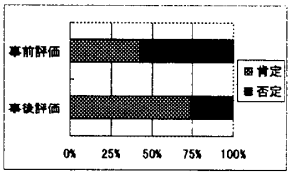

図 3　評価の変化例

図 3 の映画では，鑑賞前の評価では“見たいと思ってなかった”，“あんまり期待してなかった”といった意見の割合が多かったのに対し，鑑賞後の評価は“見てみたら意外とよかった”，“見て大正解でした”といった意見の割合が多くなっており，評価が良くなっていることが分かる．

4. 実験

4.1 概要

3 節で提案した可視化を実現するためには，各項目に対する意見を抽出し，肯定・否定の極性を判定する必要がある．本研究では，機械学習器 SVM を用いた意見抽出技術[2]により，各項目に対する肯定・否定意見の自動抽出を試みた．また，3.3 節で述べた被験者が付与したデータを正解とした時の精度・再現率により自動抽出技術の評価を行った．最後に，評価結果に基づき，意見抽出の可視化への影響について考察した．

4.2 評価

自動抽出の結果を，表 2 に示す．

表 2　自動抽出結果

	精度	再現率
監督	50.00	25.00
出演者	65.56	14.65
映像	61.39	28.27
音楽	47.64	23.15
ストーリー	49.13	18.37
全体	54.74	21.89

また，図 1，図 2 と同じ映画について自動抽出の結果に基づき可視化した例を図 4，図 5 に示す．比較すると，映画の関心の傾向を捉えていることが分かる．

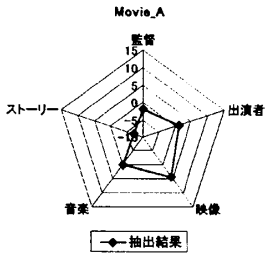

図 4 自動抽出の可視化 1

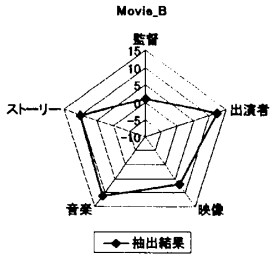

図 5 自動抽出の可視化 2

4.3 考察

本研究の目的は，映画の関心の傾向を可視化することである．項目に対する関心の傾向を捉えるためには，すべてのデータを抽出する必要はなく，特徴を示すのに十分なだけのサンプル数を抽出することが必要となる．そのため，再現率よりも精度が重要となる．

以下では，5 つの項目についての考察を行う．

- 「映像」は，分類のキーワードが正解をカバーしており，データ数も十分なことから，傾向を可視化できた．
- 「監督」は，正解データのサンプル数が少ないことから，正解をまったく抽出できなかった事例も存在した．
- 「出演者」は汎用的なキーワードを設定しているので，人名などを手がかりとした文は抽出できなかった．ただし，人名以外の手がかりを組み合わせて抽出した事例もあり，精度は良いため，傾向を示すことはできた．
- 「音楽」は，曲名等は汎用性が低いことからキーワードにできず，意見抽出の効果も薄く，精度がよくない．
- 「ストーリー」は，分類するキーワードがジャンルごとに異なる場合が多い．本研究では，汎用的なキーワードを設定していることから，キーワードによる抽出だけでは，評価のある項目を特定できない．しかし，意見抽出技術を加味すると，精度は約 5 割となった．

5. おわりに

関心の傾向を分析するために，映画の感想が書かれたブログ記事から，関心の特徴を表す項目に対する肯定・否定意見を抽出し，可視化する手法を提案した．キーワードによる項目への分類は，5 つの項目に人手で分類した情報の大半をカバーしているが，現状の意見抽出技術では，満足な再現率は得られていない．改善は今後の課題である．

謝辞

この研究の一部は，文部科学省科学研究費補助金若手研究 (B)（課題番号 18700241）を受けて遂行された．

参考文献

[1] 立石健二，福島俊一，小林のぞみ，高橋哲郎，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，”情報処理学会自然言語処理研究会，no.2004-FI-76/no.2004-NL-163，pp.1-8，Sept. 2004.

[2] Y.Seki, “Crosslingual opinion extraction from author and authority viewpoints at NTCIR-6,” Proc. of the Sixth NTCIR Workshop Meeting., pp.336-343, NII, Japan, May.2007.