ユーザの立場の違いを反映させたスポーツ記事分類手法

A Classification Method of Sports Games Articles Based on Users' Standpoints

1DS05182T 粕谷重広 Shigehiro KASUYA

1. はじめに

現在、Web 上には多数のニュース・ブログ記事が存在している. Web 上に存在する全てのニュース、ブログ記事が、ユーザの嗜好に合致する訳ではない. この原因として、ユーザが記事の内容自体に興味を持たない場合、及び記事の書かれ方がユーザの嗜好と一致していない場合が挙げられる. 前者の例としては、スポーツニュースに対して、ユーザが興味を持つ競技や試合が異なっている場合である. 一方、後者の例としては、同一の試合について書かれていたとしても想定する読者が応援するチームが異なっている場合である. 大量の記事の中からユーザの嗜好に基づいて記事を分類することは重要な問題である. 本研究では、ユーザの立場に基づいて記事を分類する手法を提案する.

2. 対象とする記事と立場

本研究では、スポーツの試合結果について書かれたニュース記事を対象とする. 立場とは、事象を判断する際の基準である. 例えば、野球では、２つのチームが対戦して勝敗が決まるが、応援するチームによって立場が異なると考えられる. この立場を考慮するとスポーツ記事は応援するチームのポジティブな面とネガティブな面、対戦相手のポジティブな面とネガティブな面という４パターンに分類できる.

3. 分類手法

分類には対象の特徴量を定義する必要がある. 藤村ら[1]の意見・評判情報の分析手法では、分類対象となる文章中の形容詞・形容動詞の出現頻度を利用している. 予備実験として、スポーツ記事の形容詞・形容動詞の出現頻度を調査した. しかし、立場によって出現頻度に差異が見られないことから、特徴量として利用するのは不適切である.

本研究では動詞に着目して、その中でも受動態の出現頻度を特徴量として利用する. 受動態は、人や物事がある動作、作用を直接受けたり、迷惑をこうむる意を表すことがある. このことから、受動態によって書かれた記事は「れる」、「られる」という表現が存在し、応援するチームにとって良くないことが起こった可能性が高いと仮定する.

4. 実験

プロ野球の記事を対象として実験を行った. 学習データとして Web 上のスポーツ報知の 2008 年プロ野球セントラルリーグクライマックスシリーズと日本シリーズの記事 81 件を利用した. それぞれの記事に含まれる動詞を抽出し、動詞数と受動態数の比を特徴量として利用した. 分類にはテンプレートマッチング法を利用した. この分類法は、クラスごとのまとまりを１つの代表ベクトルとし、代表ベクトルから近い距離のクラスに分類する手法である. 本研究では、この分類法での適合率と再現率を求めた. ここで、適合率とは、分類結果の中での正解が含まれる割合であり、再現率とは正解中で正しく分類された割合である.

本研究では、ポジティブとネガティブという基準で分類を行った.「ポジティブ」とは試合の勝敗に関係なくユーザのファンのチームの良いことを書いてある記事であり、「ネガティブ」とは逆に悪いことが書いてある記事である.

学習データに対する分類の結果を表 1 に示す. 結果から受動態の出現率によってポジティブ、ネガティブの立場への分類が高精度で実現できる可能性が示された.

次に、学習データによって得られたパラメータに対して評価データを用いた分類実験を行った. 評価データとしては中日スポーツ、日刊スポーツ、西日本スポーツの 3 紙の記事 64 件を使用した. 適合率と再現率の結果を表 2 に示している. 結果からポジティブの適合率、再現率と共に高い値が得られている. ただし、ネガティブに対する結果が低い.

表 1：適合率と再現率（学習データ）

	適合率	再現率
ポジティブ	90%	79%
ネガティブ	58%	75%

表 2：適合率と再現率（評価データ）

	適合率	再現率
ポジティブ	73%	86%
ネガティブ	60%	41%

5. チーム特定のための特徴量

ユーザの嗜好に合うものを分類するには、ポジティブ・ネガティブの分類の他にチームを選別する必要がある. 次に応援するチームを特定するために選手名の出現数での分類を行った. それぞれのスポーツ新聞によって若干違いはあったが、平均で 88%の分類精度が得られた. この結果からチームを特定するためには選手名の出現数を用いることが有効であると判断できる.

6. おわりに

本研究では、スポーツの試合結果についての記事を対象として分類に受動態の出現頻度を利用してポジティブ、ネガティブという基準での分類手法を提案した. 実験により、本手法の有効性が示された.

今後、ネガティブの分類精度を上げていくために、動詞の種類を考慮した分類が考えられる.

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，第 18 回人工知能学会全国大会，2004.