

スポーツ記事の書かれ方を利用した立場に基づく分類手法

粕谷 重広[†] 牛尼 剛聡[‡]

[†]九州大学芸術工学部 〒815-8540 福岡市南区塩原 4-9-1

[‡]九州大学大学院芸術工学研究院 〒815-8540 福岡市南区塩原 4-9-1

E-mail: [†] kasuya@gospel.aid.design.kyushu-u.ac.jp, [‡] ushiama@design.kyushu-u.ac.jp

あらまし Web 上には膨大な数のニュース・ブログ記事が存在している。大量の記事の中から、ユーザの嗜好に適した記事を選別することは重要な問題である。記事の中にはユーザが内容に興味を持たないものもあるが、書き方がユーザの嗜好に合致しない記事もある。本論文では、スポーツの試合結果について書かれた記事を対象に、記事の立場によって分類する手法を提案する。分類に考慮する立場はポジティブ、ネガティブという基準である。分類には、受動態数と動詞数の比を利用する。また、チームの判別を行うために選手名の出現数を利用する。キーワード 情報分類、立場、センチメント、レトリック

Classification of Sports Article Based on Standpoint Using Their Rhetoric

Shigehiro Kasuya[†] Taketoshi Ushiamo[‡]

[†]Department Design Kyushu University

4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan

[‡]Faculty of Design Kyushu University

4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan

Abstract There exist a lot of news and blog articles on the Web. It is one of the most important issues to select the articles that are suitable for a user. A user may not be interested in the subject of an article, and the user may not like the rhetoric of an article. In this paper, we introduce a classification method for sports articles based on user's standpoints. The standpoints to consider for classification are positive and negative. We use the ratio of the number of the passive voice and the number of the verbs for a feature value of classification. To classify articles based on the team of interest we use the number of the appearances of the player name for a feature value of an article.

Keyword Classification, Standpoint, Sentiment, Rhetoric

1. 序論

1.1 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。それらの多くの記事の中からユーザの嗜好に合った記事を発見することは重要な問題である。

Web 上に存在する全てのニュース、ブログ記事が、ユーザの要求に合致するとは限らない。記事がユーザの要求に一致しない理由として、ユーザが記事の内容自体に興味を持たない場合と、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、スポーツのニュース記事を対象とした場合、記事が対象とする試合自体にユーザの興味がない場合が挙げられる。ユーザがプロ野球の試合についての記事が読みたいとすると、スポーツニュース記事にはサッカー等他の競技のニュースもあり、野球以外のニュース記事にユーザは興味がない。一方、記事の書かれ方がユーザの施行と一致しない場合とは、試合自体には興味があるもののチームという立場が違えば、書き方も違ってくるものである。野球の試合に

関する記事で、ファンのチームの試合について書かれた記事があったとしても、相手チームの出来事ばかり書かれていたとしたら読みたいとは思わないだろう。読みたい記事はファンチームについて書かれたものである。こうした背景の下、本研究では、記事が書かれた立場に基づいて、ユーザの立場と合う記事を分類する手法を提案する。

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。対象とするユーザの立場はユーザが支持するチームである。本研究で、スポーツニュースを対象とする理由は、スポーツ新聞等によって立場が明確な記事を容易に収集することができるからである。そして、想定する読者によって記事の書かれ方が異なっているからである。スポーツ記事は基準が明確であり、分類の正解が求めやすい。

プロ野球の試合結果について書かれた記事は試合の勝敗によってユーザの勝者としての立場と敗者としての立場が生まれる。本論文では提案する分類手法を利用することで、ユーザの支持するチームの勝者とし

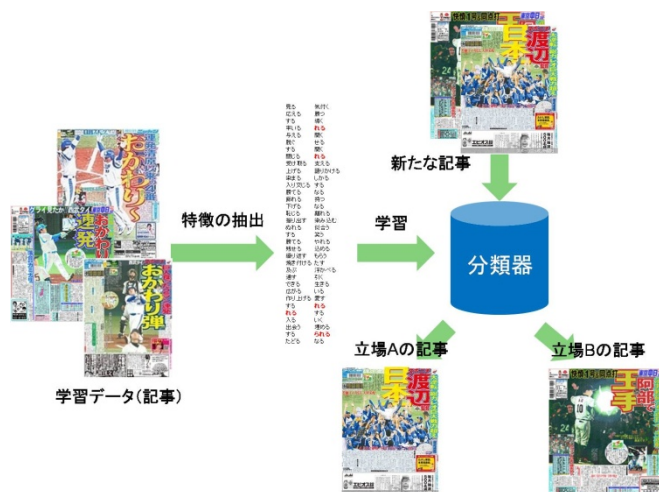


図 1：処理の概要

での立場で書かれている記事を推薦する等の応用が可能である。

提案手法では、まず Web 上に存在するスポーツ記事に対して形態素解析を行う。形態素解析した結果から文書の特徴を抽出する。抽出された特徴を分類器に学習させて、立場に基づいた記事の分類を行う。図 1 に処理の概要を示している。ここでは、予め記事を用意しておき、そこから特徴を抽出する。その特徴を利用して分類器を作成し、新たな記事の判断を行う。新たな記事は、今回は特定の試合を検索エンジンで検索した結果のページを分類器を通して推薦する。

本研究を発展させることで、ニュース記事だけでなく、ブログ記事等の分類まで適用出来ると考えられる。本研究では、対象をプロ野球だけに限定しているが、スポーツやその他のニュース記事に対しても発展していくことが可能と考えられる。

本論文の構成は以下の通りである。第 2 章で関連研究を挙げる。第 3 章で実験データとなるスポーツ新聞の記事からの分類実験の結果と分類の評価データを用いての実験結果を示す。第 4 章でチームの分類の結果を示す。第 5 章で考察を述べる。

2. 研究の位置付け

従来から、掲示板からの意見情報の抽出や抽出された意見情報を用いて意見、評判情報を判断する手法に関する研究が活発に行われている。そのような研究では、意見、評判情報を抽出するために形容詞、形容動詞が特徴語として利用されることが多い。

藤村らの研究[1]では肯定的（否定的）な評判には、肯定的（否定的）な概念を持った語が多く含まれているという仮定に基づいて形容詞、形容動詞の出現頻度から意見、評判情報を抽出する手法を提案している。この手法では、評判において特徴的な語が肯定的な評

価表現については正の値をもって、否定的な評価表現については負の値をもって抽出される。各文書に含まれる単語の属性（肯定、否定）の値の総和が 0 より大きければ、肯定的な評判であるとし、0 より小さければ、否定的な評判であるとして分類している。

情報推薦の一種として、ユーザの閲覧履歴に基づいて嗜好性の高い情報を推薦する研究もおこなわれている。河合ら[2]のニュース記事統合・推薦システム MPV（マイ・ポータブル・サイト）では、嗜好情報に基づく分類と推薦が行われている。MPV は、ユーザが普段アクセスしている Web ページのインターフェイスを嗜好情報の一つととらえ、使い慣れている既存のニュースページのレイアウトを利用し、内容のみを統合した推薦情報（カテゴリや記事）に置き換えて提示している。MPV では閲覧したページから文書の代表的な語となる興味語を $tf \cdot idf$ 値を用いて、重み付けをする。そして、興味語と閲覧履歴に基づいて記事に優先順位を付けている。MPV では興味語に関連する記事へ容易にアクセス可能となった。一方で、興味語の有無という分類基準だけでは、ユーザの好む記事と好まない記事をうまく分類できないことがあった。

上記の問題をふまえて、新たな選択指標としてセンチメント情報という分類基準を用いている。記事を分類する際に記事に出現する単語だけでなく、単語に対する書き手と読み手（ユーザ）のセンチメントという新たな選択基準を用いることで、ユーザが共感（感情移入）しやすい記事を優先的に推薦している。記事のセンチメント情報は 8 つのセンチメントを利用して、「明るい⇔暗い」、「承認⇔拒否」、「緩和⇔緊張」、「怒り⇔恐れ」という 4 つの尺度を定義して、評価値を抽出している。ここでは、あらかじめ作成している辞書から各単語の値を所得して、各尺度値を決定している。これにより、ユーザの嗜好により合った記事推薦を可能としている。

MPV ではユーザの閲覧履歴からユーザの興味のあるキーワード（興味語）を抽出し、興味語との関連度から記事を優先して推薦する。そこでは、ユーザの興味やセンチメントといった、個々のユーザの観点に合わせた情報推薦が可能になったが、一方で同一の話題が各サイトではどういった観点で書かれているかをユーザは明確にとらえることができない。そこで、ニュースサイト間の違いをセンチメントの違いとして可視化する手法として、濱砂ら[3]は、同一の話題に関して抽出したセンチメントを各サイトごとに分析することで、各ニュースサイト（新聞社）の観点の差異を情報提示できる手法を提案している。ニュースサイトごとの観点の相違を表すために、「明るい⇔暗い」、「承認⇔拒否」、「緩和⇔緊張」、「怒り⇔恐れ」の 4 つのセンチ

メント値を用いる．そして、各ニュースサイトの所在地にあわせて各センチメント値の上位 10 件の平均値を棒グラフで地図上にマッピングしてユーザに提示する．これにより、任意の球団に関してその球団の本拠地とそれ以外の地域のニュースサイトの記者がどういったセンチメントを発信しているかを一元的に解釈できるようにしている．

本研究では、従来の意見、評判情報で行う掲示板の意見が書かれた文書とは違った観点であるニュース記事の分類を対象とする．スポーツに関するニュースは出来事についての記事であり、従来の意見情報の抽出とは違った分類基準が必要となる．また、スポーツの試合結果に関するニュース記事をチームに基づいて分けること出来る点にも着目して分類する手法をとる．

嗜好情報に基づいたニュース記事の推薦で行われているセンチメントの基準は詳細に分類されているが、あらかじめ辞書を用意しておく必要がある．しかし、本研究では辞書等を用いない分類を目指している．また、従来手法では $tf \cdot idf$ 値を用いての文書中の興味語に対する重みづけを行っている．本研究では、 $tf \cdot idf$ 値を用いるのではなく、品詞の使い方の特徴に着目しての分類する．今回の研究ではニュース記事を対象とするが、ブログ等の記事にも対応するような分類手法を目標としている．

また、センチメントマップと比較した場合、立場を考慮しての差異を出す点は、本手法と同じであるが、センチメントマップでは、各ニュースサイトのセンチメント値を出してニュースサイトの観点の違いの抽出を目標としており、個々の記事の違いに注目していない．本研究では、個々の記事に対して、ユーザの嗜好に合う記事を立場から分類することを目標として研究を行っている．

3. 立場に基づく記事の分類

3.1 記事と立場

本研究では、Web 上に存在するニュース記事を対象として、それが書かれた立場に基づいて分類する手法の開発を目的とする．立場とは、事象を判断する際の基準である．ニュース記事に関する立場としては、様々なものを考えることができる．本研究では簡単化のために、野球の試合結果に関するニュース記事を対象とする．野球の試合に関する立場としても様々なものが考えられる．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、応援するチームと対戦相手のチームという2種類の立場を想定できる．立場の中で記事はファンのチームの良いこと、悪いこと、対戦相手のチームの良いこと、悪いことを書く4パターンが存在する．表1に一覧を示す．本研究では、どちらのチ

表 1：記事の書かれ方

	応援するチーム	対戦相手のチーム
良いことを書いた記事	応援するチーム良いことを書いた記事	対戦相手のチーム良いことを書いた記事
悪いことを書いた記事	応援するチーム悪いことを書いた記事	対戦相手のチーム悪いことを書いた記事

ームの立場にも立っていない中立的な立場の記事は考慮しない．

本論文では、分類学習用データとして Web 上に存在するスポーツ新聞社の記事を利用する．その理由は次の通りである．スポーツ新聞は想定している読者を、特定のチームのファンに限定していることが多い．例えば、一般にスポーツ報知は巨人の読者を想定して記事を書いていると認識されている．表2に代表的なスポーツ新聞と想定する読者が応援するチームを示す．今、特定の新聞社の記事を分類することを考える場合、応援するチームを固定できるので、応援するチームの良いことが書かれた記事か悪いことが書かれた記事かのどちらかの分類だけを考えればよい．このことにより、良いことを書いた記事、悪いことを書いた記事の分類について考えていく．

表 2：スポーツ新聞と
想定する読者が応援するチームの例

スポーツ新聞	想定する読者が応援するチーム
西日本スポーツ	ソフトバンク
スポーツ報知	巨人
中日スポーツ	中日
デイリースポーツ	阪神

3.2. 分類に利用する特徴量

3.2.1 単語の出現頻度

記事を分類するためには記事の特徴量を定義する必要がある．

従来の関連研究における意見、評判情報の分析では、特徴量として形容詞、形容動詞の出現頻度が利用されることが多い、本研究においても、形容詞、形容動詞の出現頻度を利用することが考えられる．

単語の抽出のためのシステムの実装にあたり、記事の形態素解析を行うために Slothlib[4]を介して形態素解析システム茶筌 (ChaSen) [5]を利用する．

予備実験としてプロ野球の試合結果についてのスポーツ報知の記事 45 件中に含まれる形容詞、形容動詞を抽出した。表 3 に形容詞、形容動詞の抽出結果の 1 部を示す。

表 3 の結果では各記事の出現数が少なく出現しない記事があったことと、抽出された語に差異が見られない。45 件中の文書で形態素解析を行うと、約 2 万語となったが、そのうち形容詞、形容動詞は約 200 語しかなかった。形容詞、形容動詞の出現確率は約 1%であり、出現確率が少ないために個々の記事に対する特徴量としては、利用が困難である。「いい」という形容詞は勝利の記事で出現が多いと思われる。それは、勝利の記事は、良いことが書かれていると考えられるからである。しかし、結果は敗北の記事の方が出現確率が高い。その理由として考えられるのは、応援するチームが負けて、相手チームの選手の良いことが目立ったため、「いい」という語が使われるなど必ずしも試合の最終的な勝敗だけでなく、途中経過など様々な側面に関する記述が行われるからである。今回の実験では、勝利・敗北のどちらかにおいて多く出現する、記事の特徴となり得る形容詞、形容動詞が出現しない記事が多かった。このことから特徴量として利用するのは困難であると考えた。

表 3：スポーツ記事中の形容詞、形容動詞の出現頻度

単語	勝利の記事での出現頻度	敗北の記事での出現頻度
いい	14	17
ない	8	17
悔しい	5	9
厳しい	4	3
強い	4	7
良い	4	3
大きい	4	2
悪い	1	6

3.2.2 受動態

上記の結果をふまえ、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。

例えば、2008 年度 10 月 22 日のプロ野球セントラルリーグクライマックスシリーズ第 1 戦の結果に対するスポーツ報知の記事を図 2 に示し、中日スポーツの記事を図 3 に示す。図 2 の記事では応援するチームの打者である中日の選手（中村、文中では「ノリ」と表記）に注目しており、この選手がヒットを打ったことが能動態で表現されている。一方、図 3 の記事では、応援するチームの投手である巨人の選手（クルーン）に注

中日スポーツ 試合日：10 月 22 日
ノリ絶叫V打！クルーンのフォーク粉碎Gアドバンテージ消した
決めたのはノリだった。9 回表 2 死一、三塁、3 - 3 の同点から勝ち越しの中前打を放った。

図 2：中日スポーツの記事

スポーツ報知 試合日：10 月 22 日
クルーン痛恨！中村紀に決勝打許しG初戦黒星
中村にはファウルで粘られた末に中前に運ばれ、勝ち越しを許した。

図 3：スポーツ報知の記事

目し、相手のチームである中日の選手がヒットを打ったことが受動態で表現されている。

受動態は、人や物事がある動作、作用を直接受ける。また、迷惑などをこうむる意を表すときがある[6]。そこで、受動態によって書かれた記事は応援するチームにとって良くないことが書かれている可能性が高いと考えた。この考えに基づき、受動態によって書かれていない記事は、ファンチームの良いことが書かれていると仮定する。この仮定を利用して記事の分類を行う。

4. 分類実験

4.1. 「勝利」、「敗北」の分類

上記の仮定を基に、プロ野球の試合結果に関する記事中で出現する受動態数を調査した。対象とした記事は、インターネットで公開されたスポーツ報知のクライマックスシリーズと日本シリーズの記事 81 件である。それぞれの記事に含まれる全ての動詞を抽出し、動詞数と受動態数の比を求めた。図 4 に記事に含まれる受動態数と動詞数の分布を示す。図の x 軸方向の値は動詞数、y 軸方向の値は受動態数である。丸は応援するチームが勝利した場合の記事であり、正方形は、応援するチームが敗北した記事である。受動態数と動詞数の比を特徴量として利用している。

テンプレートマッチング法を利用して分類する。ここで分類に利用するクラスは、応援するチームが勝利した「勝利」と応援するチームが敗北する「敗北」である。テンプレートマッチング法では、クラスごとのまとまりを 1 つの代表ベクトルとする。代表ベクトルの求め方は多数存在するが、クラス c に属する学習データの分布から、平均 μ_c を求めるのが最も自然かつ簡単である。新しい入力 x の識別方式としては、各クラスの代表ベクトル μ_c との間のユークリッド距離を計測して、距離が一番近いクラスへと識別する。 (x, y) をベクトル x

とyの内積として、

$$\left\|x-\mu_c\right\|^2=\left\|x\right\|^2-2\left(x,\mu_c\right)+\left\|\mu_c\right\|^2$$

なので、クラス平均までの距離（の2乗）を最小にするcを選ぶことは、クラスcごとに定義される

$$g_c(x)=2\left(x,\mu_c\right)-\left\|\mu_c\right\|^2$$

というxに関して1次（線形）の関数を最大にするcを選ぶことと同値となる。

本研究では、代表ベクトルを全動詞数と受動態数の比の勝利・敗北の平均によって求めた。本研究では、勝利の代表ベクトルの値は0.0379で、敗北の代表ベクトルの値は0.0721である。この二つの代表ベクトルから同距離にある点は直線で表せる。その直線を図に示している。直線の式は $y=0.055x$ となる。この分類法での適合率と再現率を求めた。

本研究で適合率、再現率は以下を表すものとする。

$$\text{適合率}=\frac{\text{正しく分類された記事数}}{\text{分類結果中に含まれる記事数}}$$

$$\text{再現率}=\frac{\text{正しく分類された記事数}}{\text{正解となる記事数}}$$

分類結果を表4に示す。

表4の結果から勝利での適合率・再現率は比較的高い値が得られたが、敗北に関してはどちらも50%強と十分な結果が得られていない。その原因としては、勝利した際にも調子が悪かった選手がいれば、受動態を

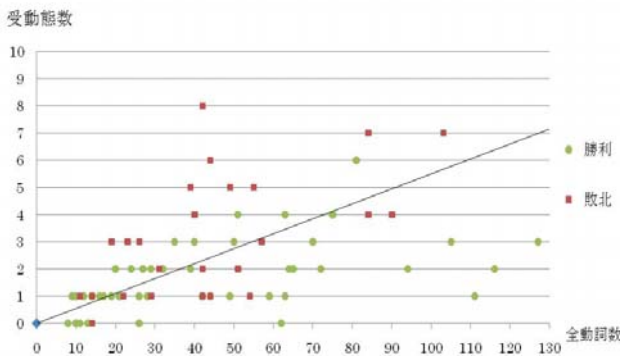


図4：受動態数と全動詞数（勝利・敗北）

表4：適合率と再現率（勝利・敗北）

	適合率	再現率
勝利	74.4%	68.1%
敗北	51.6%	59.3%
全体	64.9%	64.9%

3 回降板の上原「もう少し投げないと」...

日本S第5戦

初回、先頭の片岡のバットをへし折った打球が、右前に抜けた。後続に連打さ~~れ~~、無死満塁のピンチを招くと、1死後に石井義の二ゴロ併殺崩れの間に先取点を奪わ~~れ~~た。

図5：勝利記事での受動態が使用される例

利用して書かれる等、全てが良いことが書かれている記事とは限らないからである。例を挙げると、2008年11月6日の日本シリーズ第3戦のスポーツ報知の記事では、巨人が勝利したにも関わらず図5のような記事が書かれていた。この記事では、勝利した試合の中でも、期待を裏切った選手のことについても書かれている。勝利、敗北での分類が上手いかわからない理由の一つは、こうしたことにありと考えられる。そこで、勝利した試合での悪いことを書いた記事、敗北した試合での良いことを書いた記事を考慮した分類基準であれば、高精度の分類が期待できると考えた。

4.2.「ポジティブ」、「ネガティブ」への分類

上記をふまえ「勝利と敗北」ではなく、「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が応援するチームに関して良い出来事について書いてある記事であり、「ネガティブ」とは逆に応援するチームの悪い出来事が書いてある記事である。

「勝利」、「敗北」と同様にスポーツ報知の記事81件を利用して、分類を行った結果を図6に示す。適合率・再現率の結果を表5に示す。ここで、予め正解となるポジティブ、ネガティブの記事の判断は著者が主観的にで行った。

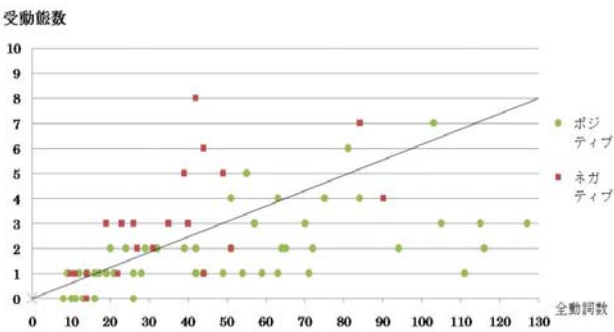


図6：受動態数と全動詞数(ポジティブ・ネガティブ)

表 5：適合率と再現率（ポジティブ・ネガティブ）

	適合率	再現率
ポジティブ	89.8%	78.6%
ネガティブ	55.6%	75.0%
全体	77.6%	77.6%

表 5 の結果から「勝利」、「敗北」の分類より「ポジティブ、ネガティブ」の分類が適合率・再現率と共に上昇していることが分かる。この結果から、受動態の出現率によって「勝利」、「敗北」よりも「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。

ユーザにとってファンチームの良い出来事を書いた記事とは、ファンチームの勝利の記事とはかぎらない。ファンチームの悪い出来事を書いた記事とは、ファンチームの敗北の記事とは限らない。勝利、敗北に関係なく良い出来事、悪い出来事を書いた記事が存在する。勝利、敗北という基準の分類では、ユーザにファンチームの良いことを書いた記事を推薦するというシステムであると、勝利の記事だけがユーザに推薦され、敗北の時は、何も推薦されないという問題が生じてしまう。しかし、勝敗に関係なく分類できることにより、ファンチームの敗北の際にもユーザに良い出来事を書いた記事の分類を行うというユーザの嗜好により合った分類手法が提案できる。

3.4.評価データの分類

次に 4.1.で得られた分類パラメータを利用して、他の新聞社のデータを評価データとして実験を行った。用いた記事は中日スポーツ、デイリースポーツ、西日本スポーツの 3 社の記事 64 件である。

結果の分布図を図 7 に適合率と再現率を表 6 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再現率の結果が他に比較して低い。

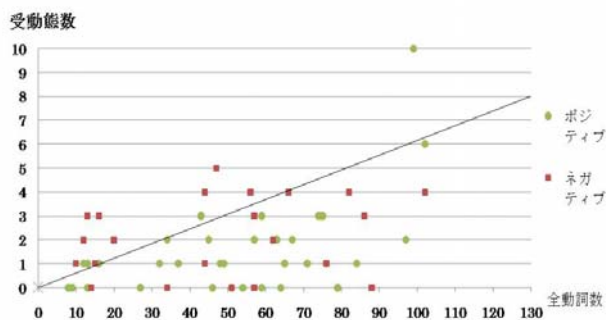


図 7：受動態数と全動詞数（評価データ）

表 6：適合率と再現率（評価データ）

	適合率	再現率
ポジティブ	73.5%	85.7%
ネガティブ	60.0%	40.9%
全体	70.3%	70.3%

次に評価データでは学習データとは異なり、複数の新聞社の記事を利用したので、新聞社ごとの適合率と再現率を求め、差異があるかを示す。西日本スポーツの記事 27 件を表 7 に、中日スポーツの記事 24 件の結果を表 8 に、デイリースポーツの記事 13 件の結果を表 9 に示す。

各新聞社によって違いはあるものの「ポジティブ」に関する再現率、適合率は高い値が得られている。西日本スポーツと中日スポーツでのネガティブの適合率、再現率は分類に利用可能な値が得られているが、デイリースポーツでは、ネガティブの再現率が低くなっている。これは、デイリースポーツの記事数が少なく、ポジティブ、ネガティブの再現率が 100%であったが、ネガティブと分類された記事数が少ないことでこのような結果となった。また、ネガティブと分類された記事数が少ないのは、負けた試合で受動態ではない言葉でのネガティブな表現が多かったためと考えられる。

表 7：適合率と再現率（西日本スポーツ）

	適合率	再現率
ポジティブ	66.7%	75.0%
ネガティブ	55.6%	45.5%
全体	63.0%	63.0%

表 8：適合率と再現率（中日スポーツ）

	適合率	再現率
ポジティブ	87.5%	77.8%
ネガティブ	50.0%	66.7%
全体	75.0%	75.0%

表 9：適合率と再現率（デイリースポーツ）

	適合率	再現率
ポジティブ	66.7%	100%
ネガティブ	100%	20.0%
全体	69.2%	69.2%

5. 応援するチームの特定

前章まで、受動態の出現率による「ポジティブ」、「ネガティブ」の二つの立場に分類をしてきた。しかし、これだけではユーザに推薦する記事を特定することができない。その理由は、それぞれの記事がどのチームを応援する読み手を想定して書かれているかが不明だ

からである．そこで、ユーザが応援するチームのことが書かれた記事と対戦相手のチームのことが書かれた記事を分類するためにチームを特定する手法について考える．

チームを特定するための特徴量として、選手名の出現頻度を利用する．例えば、巨人対中日の試合であれば、巨人を応援する読者を想定している記事は、巨人の選手名が多く出現するであろうし、中日を応援する読者を想定している記事は中日の選手が多く含まれると予想されるからである．これは、ユーザは応援するチームについて書かれた記事を読みたいはずであり、ファンチームの選手名が多いことは、そのチームについて書かれていると仮定する．

対象とする記事として、スポーツ報知、中日スポーツ、日刊スポーツ、西日本スポーツの記事、計 149 件を用いて調査を行った．スポーツ報知は巨人の選手名、中日スポーツは中日の選手名、日刊スポーツは阪神の選手名、西日本スポーツはソフトバンクの選手名とそれぞれのチームの選手が多く出現すれば正しく分類されたとする．結果を表 10 に示す．

この結果から、新聞によって多少の差があるが、十分に利用できるものである．これより、チーム特定のための特徴量としては選手名の出現頻度が有効であることが示された．

本手法を実現する際には、選手名が上手く抽出できない場合がある．具体的には、外国人選手の名前がカタカナで表現されると一通りとは限らなく、正しく解析ができない．この問題に対しては、全ての選手名を予め辞書に登録しておくことで解決することができる．しかし、例えば、中日の「中村選手」を「ノリ」と表現されているなどの問題は対処できない．この問題を解決するために、外間ら[7]は、一般的な Web 文書では、あるオブジェクトは公的・正式な名称だけでなく様々な呼び名で参照される、という点に着目し、人物の呼称を抽出することを目的として、人物のフルネームが出現するパターンを手がかりに、人物の呼称を抽出する手法を提案している．この手法では、「こと **fullname**」という文字列の直前に出現する文字列からの呼称候補を抽出することとフルネームと隣接する文字列を抽出して、フルネームとパターンの関連の強さ

を基に重みづけを行い、重要なパターンを「隣接パターン」として選択する．隣接パターンと呼称候補を評価し、評価値の高い候補上位を呼称として出力している．

このように、呼称や愛称を抽出するための方法を用いることでこの問題は解決できると期待できる．

6. おわりに

本論文では、大量のニュース、ブログ記事の中から立場に基づいたニュース、ブログ記事の推薦に利用するための分類手法を提案した．スポーツの試合結果についての記事を対象として、受動態を利用して「勝利」、「敗北」での分類を試みた．さらに、この分類結果をふまえ、「ポジティブ」、「ネガティブ」という基準での分類を行った．「勝利」、「敗北」の基準での分類より「ポジティブ」「ネガティブ」の方が精度が上がり、高精度で分類可能であるという結果が示された．また、チームの特定のために選手名の出現頻度を利用したの分類が十分に利用できることが明らかになった．

今回の実験では、「ポジティブ」、「ネガティブ」で分類を行ったが、受動態の分類での評価データではネガティブの分類精度が低かった．その原因としては、主語、述語の関係を考えなかったからであると考えられる．受動態が必ずしもユーザの応援するチームだけに使われるとは限らない．他には、受動態が使われなくても応援するチームの悪いことが書かれることを考慮していないので、このような結果になったと考えられる．この問題を解決するためには、「ポジティブ」、「ネガティブ」の分類における特徴となり得る他の特徴を見つけていく必要がある．この目的として、動詞の使われ方には特徴があるように思われる．例えば、「ホームランを浴びる」という表現は「ネガティブ」なできごとであり、「ポジティブ」、「ネガティブ」で出現する動詞の辞書を作成することも可能である．本研究では、単語の辞書の作成には時間の問題や個人の判断だけでは行うことが出来ないと思われたので、行わなかったが、精度の向上には必要であると考えられる．

本論文では、受動態の出現頻度を利用したの分類とチーム特定のための分類を行った．しかし、本研究では、記事のパターンを 4 つ存在すると考えているので、4 つのパターンにおける分類を行えるような分類手法を考え、実験を行っていきたい．

また、本研究では、プロ野球の試合結果の記事を分類対象と考え記事を収集してきたが、実験を行うには記事数が少なかったので、少ない記事でいかに分類できるかを考慮して研究を行ってきた．記事数を増やすことで、分類の手法が異なってくるとも考えられる．

本研究の手法の発展として考えられるのは、ニュー

表 10：チームの分類

スポーツ新聞	正解率
スポーツ報知	89.0%
中日スポーツ	92.0%
デイリースポーツ	92.3%
西日本スポーツ	75.9%
平均	87.2%

ス、ブログ検索においての嗜好に基づいた記事のリランキンングやフィリタリッングがある．ユーザが読みたいチームと「ポジティブ」、「ネガティブ」度合を指定してニュース結果のリランキンングに応用できると考えられる．さらに、プロ野球以外のスポーツ競技への適応が可能と考えられる．しかし、本研究では、チームという立場を考慮していたので、他のスポーツ競技への発展は1チーム対1チームで行われる試合のみとなってしまう．そこで、個人選手や複数のチーム、グループでの立場の分類を行う必要がある．

参考文献・URL

- [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.
- [2] 河合由起子，嗜好情報に基づくニュースコンテンツの推薦とその応用 - 画一的な便利さと多様な嗜好への適応 - ，社団法人情報処理学会，Vol.48, No.9, pp. 979-983, Sept2007
- [3] 濱砂佳貴，河合由起子，熊本忠彦，田中克己，センチメントマップによる複数ニュースサイトの差異情報可視化手法の提案，DEWS2008，2008
- [4] slothlib，<http://www.dl.kuis.kyoto-u.ac.jp/slothlib/>, 2007
- [5] 奈良先端科学技術大学院大学松本研究室，“日本語形態素解析システム茶筌 Version 2.3.3”，2003，<http://chasen.naist.jp/wiki/ChaSen/>
- [6] 日本語大辞典
- [7] 外間 智子，北川 博之，Webデータを用いた人物の呼称抽出，電子情報通信学会技術研究報告，Vol.106, No.149, pp. 113-118, July2006