学士学位論文

Webニュースを利用した活躍度の日米間比較

茨城大学工学部
システム工学科
執筆者： 大北高広

指導教官： 新納浩幸

平成17年3月2日

目次

図目次

表目次

第1章　はじめに

1.1　研究概要

現在、インターネットの普及とともに大量の情報がWeb上に発信されている。その中でも、評判情報に注目が集まってきている。評判情報とは自由に意見を述べたものであり、Web上で評判情報を検索することが一般化してきている。注目が集まることで、Webの社会的影響力は加速度的に増大してきている。その為、Webを用いた評判情報分析を行うことの重要性が高まってきている。企業においては自社製品のマーケティングやクレーム処理の支援に、また個人においては商品購入等の意志決定支援などに用いることができる為である。このような、有益な情報を取り出すことができるので、評判情報を自動的に抽出することに対する期待が高まってきている [1]。

例えば、携帯電話を買い換えようしていた際、自分が欲しい携帯電話についてwebを利用して評判を調べると、「デザインが良い」と肯定的感じている人もいれば、「ボタンが押しにくい」と否定的に感じている人もいる。評判情報は使用した人、見た人により感じ方はさまざまに書かれている。このような情報がWebにはたくさん書かれている。これらのWebの情報より自分にとって必要な情報を取り出していくことが必要である。「デザイン」や「使いやすさ」など自分が何に重点を置いているかにより、この携帯電話は買うのが良いのか、違う物が良いのかと判断する材料にすることができる。このように意思決定支援として使用することができる。しかし、評判情報は様々な意見を様々に書いたものであり、書いた本人によっても評判は変わってくるのは当然である。したがって、私達が普段目にしているWebニュース記事も書いた人により評価は変わっているはずである。現在アメリカメジャーリーグにおいてたくさんの日本人選手が活躍しているが、これは日本の色々なWebニュース記事において大々的に報道されている。しかし、日本人選手の活躍が地元のアメリカ人の目にはどのように映っているかは、アメリカのWebニュース記事を読む必要がある。しかし、実際に英語の文章を読むこと

は大変なことである。

そこで本研究では、上記のように重要性が高まってきている評判情報の中でも、同一出来事に対する日本とアメリカのWebニュース記事の評判情報に、どのような差位があるのかの分析する。特に日本人メジャーリーガーについて書かれている日本とアメリカの異なるWebニュース記事に注目し、評判を日本人選手の活躍に置き換えることとする。そして日本とアメリカの異なるWebニュース記事より、それぞれ活躍情報の度合(以下、活躍度と呼ぶ)を数字として算出し比較を行っていく。その際、肯定・否定の分類も行う。

1.2 本論文の構成

本論文は、従来の手法について(第2章)、活躍度の比較方法について(第3章)、Naive Bayes法を用いた活躍度の算出方法について(第4章)、実験および結果(第5章)、その考察(第6章)からまとめ(第7章)へと進む。

また、巻末にはプログラムのソースリストとNaive Bayes法により算出した実験結果を添付した。

第2章　従来の手法

2.1　マッピングによる評判情報の分類

文章を形態素解析し、その結果をあらかじめ作成しておいた評価表現辞書と比較することにより、良い評価をしている表現数(以下Good)と悪い評価をしている表現数(以下Bad)のカウントを行う。その後、X軸をGood、Y軸をBadとすることでXY平面へマッピングする。グラフ化の結果(図2.1)、Good、Badともに多いのは詳しく評価しているリッチであり、Goodが多いものはポジティブと考えることができる。グラフ化することにより特性を容易に判断することができる[2]。

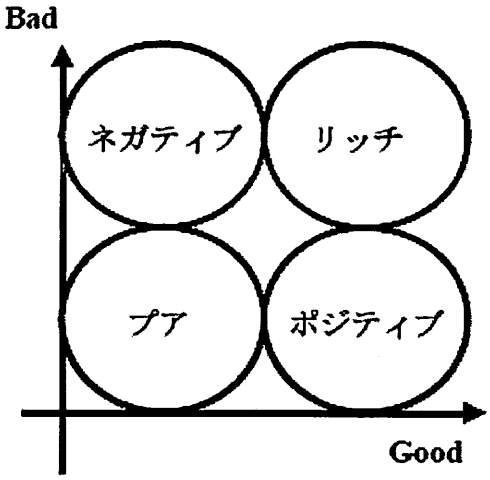

図2.1: マッピング.

このマッピングを用いた例として、Webページの内容を評価しこれをマッピングする。マッピングした点をクリックするとWebページが新しいウィンドウできるようにする。これは図2.2のように表わすことができる。このようにWebページをマッピングすることで、見る価値のある自分にあった特性のあるWebページを容易に判定することが可能となる。例からも分かるように、マッピングするこ

とにより可視化され特性を瞬時に判断し、分類することができる。

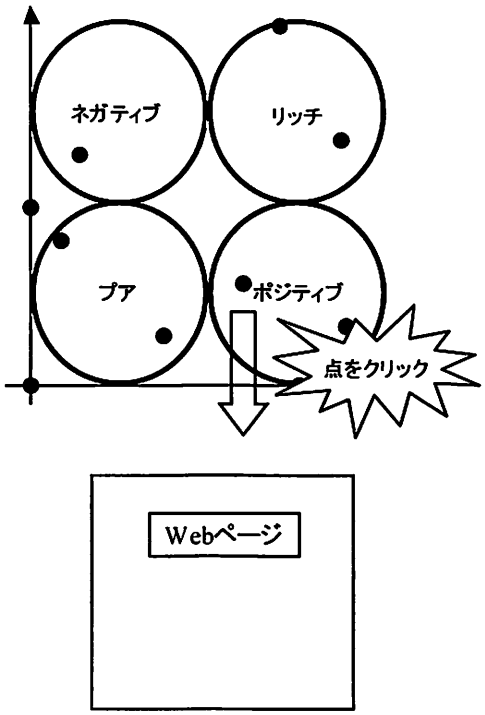

図 2.2: マッピングの例.

2.2 PN 分類手法

評価表現を抽出し評判の肯定・否定の分類(以下、PN 分類と呼ぶ) を行う。PN 分類を行うためには、まず評判で用いられる特徴的な「語」である評価表現で肯定・否定どちらの表現であるかまで記された評価表現の集合である評価表現辞書を作成する。そして言語資料体、個別言語・一作家のテキストや発話を大規模または網羅的に集めたものであるコーパスとなる対象を選ぶ。これを訓練コーパスと呼ぶ。次に属性の選択である。例として、形容詞、形容動詞のみを属性とする。これは、主に日本語でモノの評価を表す表現であるからである。そしてスコアリングの手法である。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

多く含まれているはずである。この仮定を元に、肯定的な評判と否定的な評判の差をとる。一般的な語はどちらの文書にも同様に出現するはずであるから、その影響は打ち消される。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。実際には、次の式でスコアリングを行っている。

$$score(w_i) = \frac{P_P(w_i) - P_N(w_i)}{P_P(w_i) + P_N(w_i) + k}$$

$$(-1 \leq score(w_i) \leq 1)$$

ここで、$P_P(w_i)$ は肯定的な評判で属性 w_i が出現する確率である。同様に $P_N(w_i)$ は否定的な評判である。また k は、例えば $P_N(w_i)$ が 0 であった際に、$P_P(w_i)$ が 0.1 でも 0.8 でも結果としてスコアが 1 となってしまうという、$\frac{1}{1}$ の問題を解決するために分母に加えた実数である。

ここで紹介した PN 分類法について下式に示す。

$$Score(d) = \sum_{ALL_{w_i}} score(w_i)$$

$$\begin{cases} if. \\ Score(d) > 0 \rightarrow positive \\ Score(d) < 0 \rightarrow negative \end{cases}$$

各文書に含まれる属性のスコアの総和が 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いうように分類する [3]。

上式は評判を PN に分類するので、評判の良し悪しを決定するルールそのものである。したがってその分析を行うことにより、現在のトレンドや潜在的なニーズを掴むことができる。この PN 分類手法は従来から用いられている機械学習手法と比較してもほぼ同程度の精度を得ることができる。名詞を属性に入れる際に形容詞・形容動詞と同格に扱うのではなく、何らかの工夫を行うことによってさらなる精度向上が期待できる。

第3章　活躍度の比較方法

日本とアメリカのWebニュース記事の活躍度を用いて比較する方法を説明する。

現在Webニュース記事を見ていれば、日本の出来事だけでなく、世界の出来事も大きな出来事であれば載っている。しかし、日本のWebニュース記事などを見ていると日本を中心に内容が書かれており、世界の他の人の目にはどのように映っているのか分からない。これは政治、経済、だけではなくスポーツの分野でも同じである。スポーツにおいて、日本のWebニュース記事を読んでいて日本人選手が活躍していることが大々的に載っている。それは、日本人選手について私達が興味を持っていて知りたいと思っていることであり、当然のことである。サッカーにおいて、「点を取る起点になった」「良い動きをしていた」のように書かれているのをよく目にする。メジャーリーグについては「貴重な追加点を叩き出した」「勝利に貢献した」など大活躍をしているように書かれている。しかしこれらは日本人から見た評判であり、他の国の現地の人達には実際にはどのように映っているのだろうと考えた。そこで今回は日本のWebニュース記事とアメリカのWebニュース記事との比較を行う。比較を行うことにより日本での評判情報とアメリカでの評判情報には違いがあるのかが分かると考えた。

研究で使用するWebニュース記事のサイトを日本とアメリカともにあらかじめ決めておく。そこで日本での情報収集はスポニチアネックス(http://www.sponichi.co.jp/)のWebニュース記事を使用することにした。アメリカでの情報収集はNEW YORK POST(http://www.nypost.com/) のWebニュース記事を使用することにした。両方のWebニュース記事は毎日更新され実験に使用するため、数多くのデータを収集することができると考えたので利用することにした。比較する対象として、今現在たくさんの日本人選手活躍しているアメリカメジャーリーグの記事を扱うことにした。日本人がたくさん活躍しているため、アメリカ人だけではなく、多くの日本人がメジャーリーグに注目しているためである。日本で注目が集まっている

が、実際アメリカではどのようになっているかという点からメジャーリーグのWebニュース記事を比較するのが良いと考えた。スポニチアネックスの記事の例を図3.1に、NEW YORK POSTの記事の例を図3.2に示す。

図3.1: スポニチアネックスの記事の紹介.

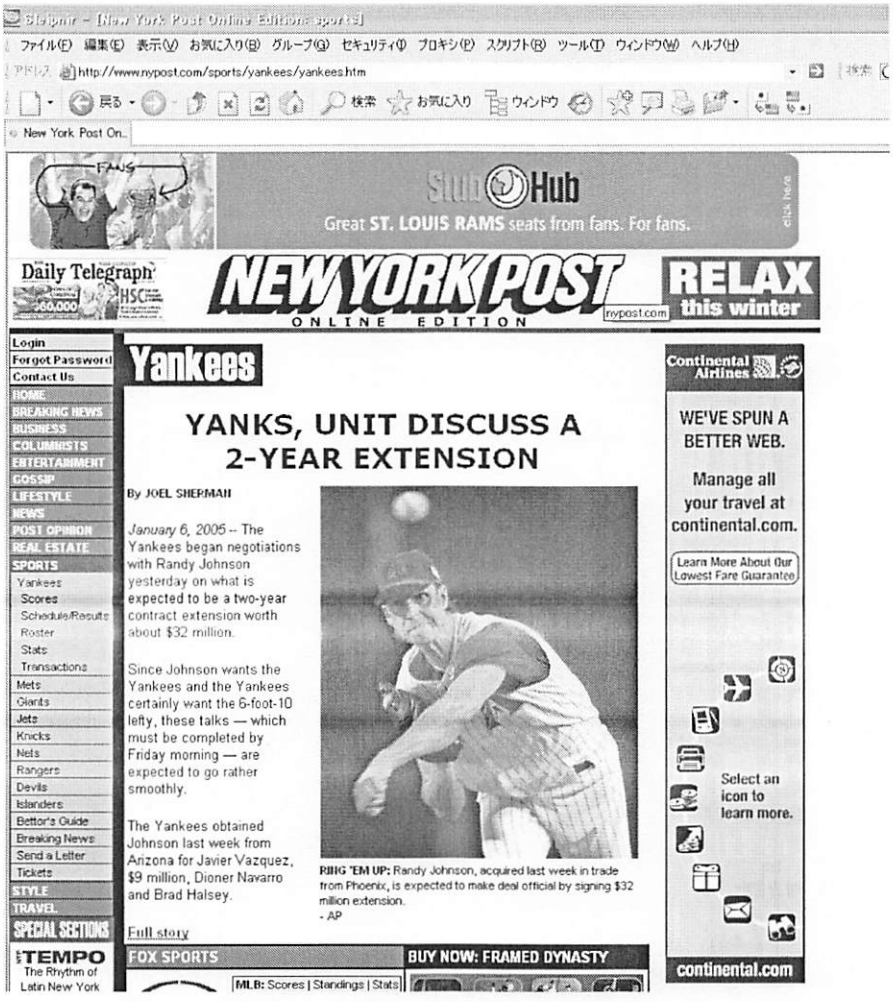

図 3.2: NEW YORK POST の記事の紹介.

本研究では、メジャーリーグで活躍する日本人選手の中でも松井秀樹選手(図3.3)に注目した。松井選手はニューヨークヤンキースというアメリカでも1番人気のあるチームに所属しており、1年目から全試合出場し大活躍をしている。日本にいる時は巨人の四番打者であり、日本人を代表するホームランバッターである。

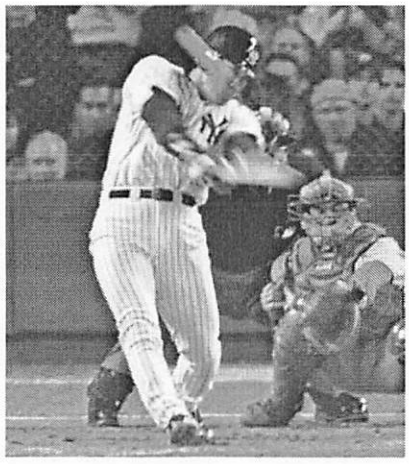

図 3.3: 松井選手の写真.

さらに、松井選手は野手であるため毎日試合に出場するので、投手より記事が豊富にありデータを取り易いと考えた。比較する際、Webニュース記事の評判情報を松井選手の活躍に置き換えることにする。スポーツ選手の記事を扱うので評判と活躍とはあまりかわるものではないが評判より活躍の方が合っていると考えたからである。日本とアメリカのWebニュース記事よりそれぞれ活躍度を算出する。活躍度を定量化し、○、×の分類も行う。そしてWebニュース記事がどの程度活躍したものであるかを算出し、日本とアメリカでは活躍の評価が異なるか否かの比較を行う。

活躍度の比較の流れを図 3.4 に示す。

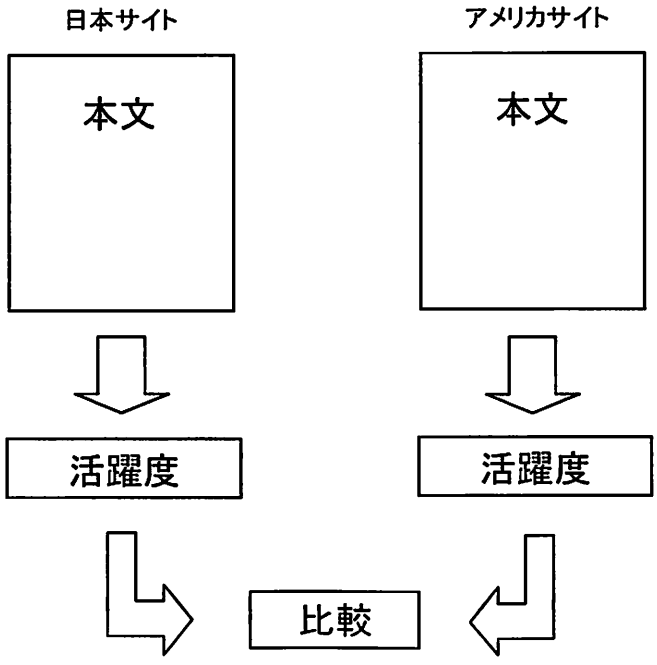

図 3.4: 活躍度の比較法.

第4章　活躍度の算出方法

4.1　Naive Bayes法

Webニュース記事より活躍度を算出するにあたり本研究ではNaive Bayes法を用いる。Naive Bayes法は自然言語処理の個々の問題を分類問題として定式化し、帰納学習の手法により解決するというアプローチである。そこで利用されてる帰納学習手法は主に確率統計学的な手法である。特にNaive Bayes法は単純なモデルであるにもかかわらず、比較的よい規則を学習できるために広く利用されている[4]。

ベイズの定理は、Thomas Bayesによって提唱された確率理論である。何かが起こる可能性はその事柄の過去の発生頻度を使ってほぼ推測できるというものである。現在では、サーチエンジン大手のGoogleと情報検索ツールを販売するAutonomyの両社もベイズの原理を採用し、高い確率で適当なデータを探し当てる検索サービスを提供している。様々な分野の研究者も、特定の症状と病気の関連付けや個人用ロボットの創造、過去のデータや経験に基づく指示に沿って行動し「考える」ことができる人工知能デバイスの開発などにベイズモデルを使っている。

ある事例xが素性f_iのリストとして、以下のように表現されたとする。

$$x = (f_1, f_2, ...f_n)$$

xの分類先のクラスの集合を$C = c_1, c_2, ..., c_n$と置く。分類問題は$P(c|x)$の分類を推定することで解決できる。実際に、xのクラスc_xは以下の式で求まる。

$$c_x = arg\max_{c \in C} P(c|x)$$

ベイズの定理を用いると、

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} \tag{4.1}$$

なので、結局式(4.2)が成立する。

$$c_x = arg\max_{c \in C} P(c)P(x|c) \tag{4.2}$$

ここで、$P(c)$ は比較的簡単に推定できる。問題は $P(x|c)$ の推定であるが、この推定は現実的に難しい。そこで Naive Bayes 型のモデルは、素性間の独立性を仮定して、以下の変形により $P(x|c)$ の推定を行う。

$$P(x|c) = \prod_{i=1}^{n} P(f_i|c) \tag{4.3}$$

結局 Naive Bayes 法の学習とは $P(c)$ と $P(f_i|c)$ の推定である。これらの値はどちらも訓練データから得られるクラスの頻度や素性の単語とクラスの頻度より推定できる。

$$P(c) = \frac{\text{訓練データ中のクラス } c \text{ のデータ数}}{\text{訓練データの数}} \tag{4.4}$$

$$P(f_i|c) = \frac{\text{訓練データ中のクラス } c \text{ の「} f_i \text{」の数}}{\text{訓練データ中のクラス } c \text{ の単語の数}} \tag{4.5}$$

4.2 Naive Bayes法の使用例

例えば、文章がA,B,C,D,E,Fと6つあったとする。これらを訓練データA,B,C,D,E,Fとし、分類したいクラスを$C = \{c_1, c_2, c_3\}$「水戸」,「日立」,「東海」と3つにしたい。ここで分類したいクラスとは、文章が水戸についてを説明している、または日立について説明しているというように分けることである。そこで文章A,B,Cが「水戸」、文章D,Eが「日立」、文章Fが「東海」となるとする(図4.1)。

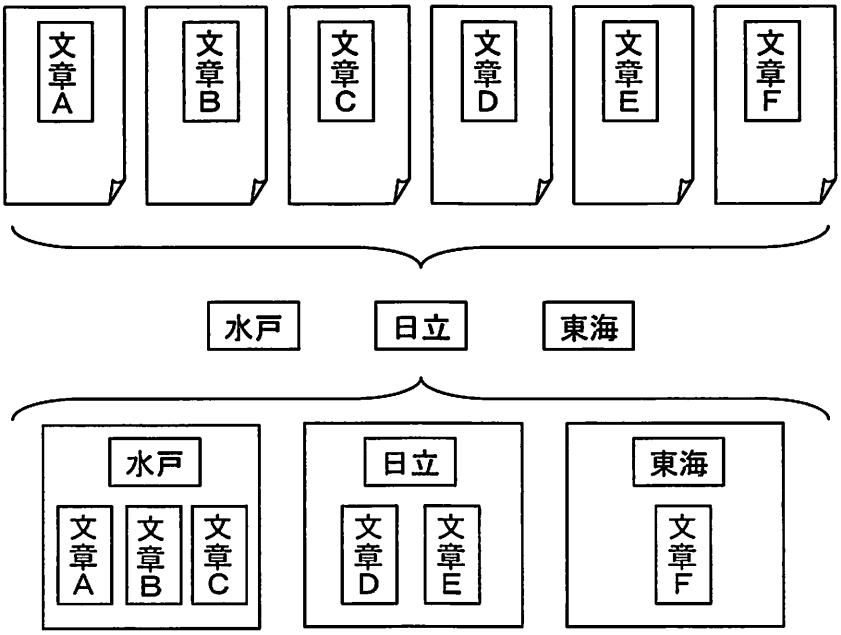

図4.1: 文章の分類例.

この時、訓練データの文中に、f_i =「茨城」という単語が「水戸」というクラス(文章A,B,C中)には10回、「日立」のクラス(文章D,E中)には6回、「東海」のクラス(文章F中)には2回出現したとする。このようにクラス中に出現する同一単語ごとにまとめ、その出現回数をカウントする。訓練データの文中に出現するその他の単語においても同様にカウントする。この様に、訓練データに出現する単語をクラスごとにまとめ、出現回数を数え上げ表にまとめたものを頻度分布表という表5.2。

表 4.1: 訓練データの頻度分布表.

単語名	クラス「水戸」	クラス「日立」	クラス「東海」
茨城	10	6	2
納豆	10	2	1
…	.	.	.
…	.	.	.
…	.	.	.
…	.	.	.
単語の数	100	60	20

Naive Bayes 法の学習は $P(c)$ と $P(f_i|c)$ の推定であるので、まずこれらの値を算出する。例えば式 (4.4) と式 (4.5) を用いて P(水戸) と P(茨城 | 水戸) を算出してみる。

$$P(\text{水戸}) = \frac{\text{訓練データのうちクラス「水戸」のデータ数}}{\text{訓練データの数}}$$

$$= \frac{3}{6}$$

$$= \frac{1}{2}$$

$$P(\text{茨城} \mid \text{水戸}) = \frac{\text{訓練データのうちクラス「水戸」の単語『茨城』の数}}{\text{訓練データのうちクラス「水戸」の単語の数}}$$

$$= \frac{10}{100}$$

$$= \frac{1}{10}$$

P(水戸) と P(茨城 | 水戸) 以外にすべての $P(c)$ と $P(f_i|c)$ を算出しておく必要がある。P(水戸) と P(茨城 | 水戸) を算出した時と同様に、式 (4.4) と式 (4.5) を用いて算出する。すべての $P(c)$、$P(f_i|c)$ を求めた結果を表 4.2、表 4.3 に表す。

表 4.2: $P(c)$ の値.

P(水戸)	P(日立)	P(東海)
$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

表 4.3: $P(f_i|c)$ の値.

P(茨城 | 水戸)	$\frac{1}{10}$	P(茨城 | 日立)	$\frac{1}{10}$	P(茨城 | 東海)	$\frac{1}{10}$
P(納豆 | 水戸)	$\frac{1}{10}$	P(納豆 | 日立)	$\frac{1}{30}$	P(納豆 | 東海)	$\frac{1}{20}$
...	.	...	.	...	.
...	.	...	.	...	.
...	.	...	.	...	.

ここまでは頻度分布表や $P(c)$、$P(f_i|c)$ のように訓練データを用いて算出する手順を述べ、Naive Bayes 法を使用する準備を行った。次に、Naive Bayes 法を実際に適用する。以下のような文があったとする。

「茨城の特産物は納豆だ。」…… (*)

この文が「水戸」「日立」「東海」のどのクラスに属するのかを分類する。まず形態素解析を行って単語ごとに分割する。

「 茨城 / の / 特産物 / は / 納豆 / だ / 。 」

分割した文は $x = (f_1, f_2, f_3, f_4, f_5, f_6)$ と表される。そして、先ほどの訓練データの単語の中でこの文の単語と一致した単語は「茨城」と「納豆」の2つとする。このとき式 (4.3) を用いて、$P(x|$ 水戸$)$ を求める。しかし、このまま計算を行うと、一致しなかった単語の $P(f_i|c)$ は 0 なので $P(x|c)$ はすべて 0 となってしまう。これを回避するために一致しなかった単語の $P(f_i|c)$ を微小な数として扱うことにした。

$$P(x|\text{水戸}) = \frac{1}{10} \times \frac{1}{10}$$
$$= \frac{1}{100}$$

同様に、$P(x|$ 日立$)P(x|$ 東海$)$ を求める。

$$P(x|\text{日立}) = \frac{1}{300}$$

$$P(x|\text{東海}) = \frac{1}{200}$$

次に、式(4.1)を用いて$P(c|x)$を求める。ただし$P(x)$はクラスcによらない値であるため、式(4.2)によるクラス判定には影響しない。したがってここではその値を求めることはしない。

$$P(\text{水戸}|x) = \frac{1}{2} \times \frac{1}{100}$$

$$= \frac{1}{200}$$

同様に$P(\text{日立}|x)$、$P(\text{東海}|x)$を求める。

$$P(\text{日立}|x) = \frac{1}{900}$$

$$P(\text{東海}|x) = \frac{1}{1200}$$

文章(*)がどのクラスに属するかは式(4.2)で決定する。3つクラスの$P(c|x)$を比べるとが一番大きい値は$P(\text{水戸}|x)$であるので、先ほどの文章(*)はクラス「水戸」に属すると推定される。このようにNaive Bayes法を用いて、与えられた文章がどのクラスに属するのかを分類していくことができる。

4.3 活躍度の算出手順

Web ニュース記事より活躍度を算出する方法を説明する。活躍度の算出の大まかな流れを図 4.2 に示す。

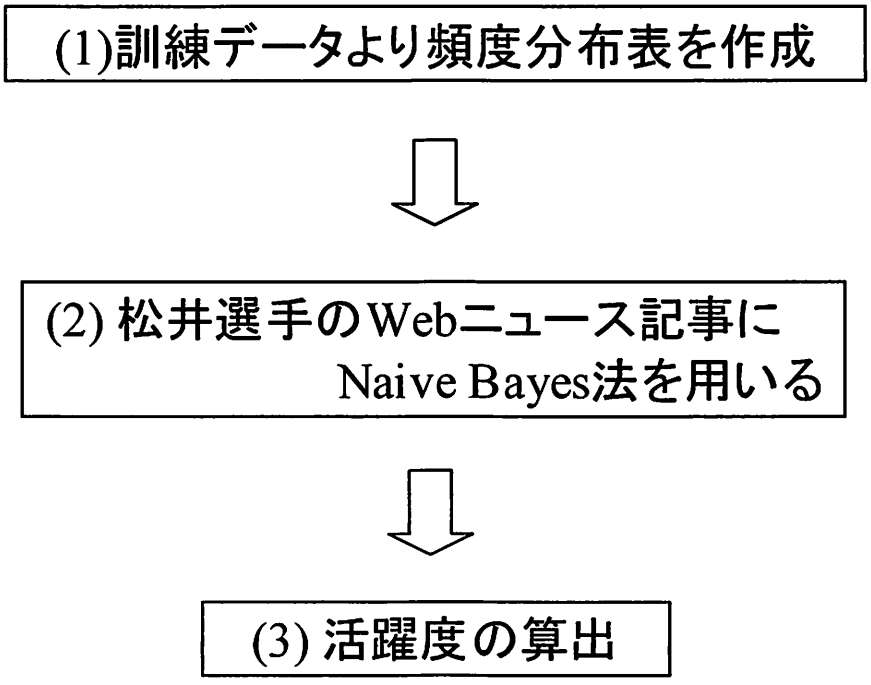

図 4.2: 活躍度の算出の大まかな流れ.

(1) 訓練データより頻度分布表を作成する。

Step1. 松井選手が明らかに活躍したもしくは活躍できなかったと分かる Web ニュース記事を集める。その際、同じような内容よりも異なる内容のものを選ぶ方が、 Naive Bayes 法を使用する際に $P(f_i|c)$ となる単語が多くなるため良い。

Step2. 活躍できた記事と活躍できなかった記事として集めたWebニュース記事から、それぞれ本文をテキストファイルとして抽出する。作業を行っていく上でテキストファイルとして抽出することは扱い易い。

Step3. 抽出したテキストファイルを形態素解析する。形態素解析とは分かり易く述べると品詞ごとに文章を分割していくことである。

Step4. 形態素解析を行ったものから名詞を抽出する。

Step5. 取り出した名詞から頻度分布表を作成する。Naive Bayes法で使用しやすいように表5.2のようにまとめる。

訓練データより頻度分布表作成の流れを図4.3に示す。

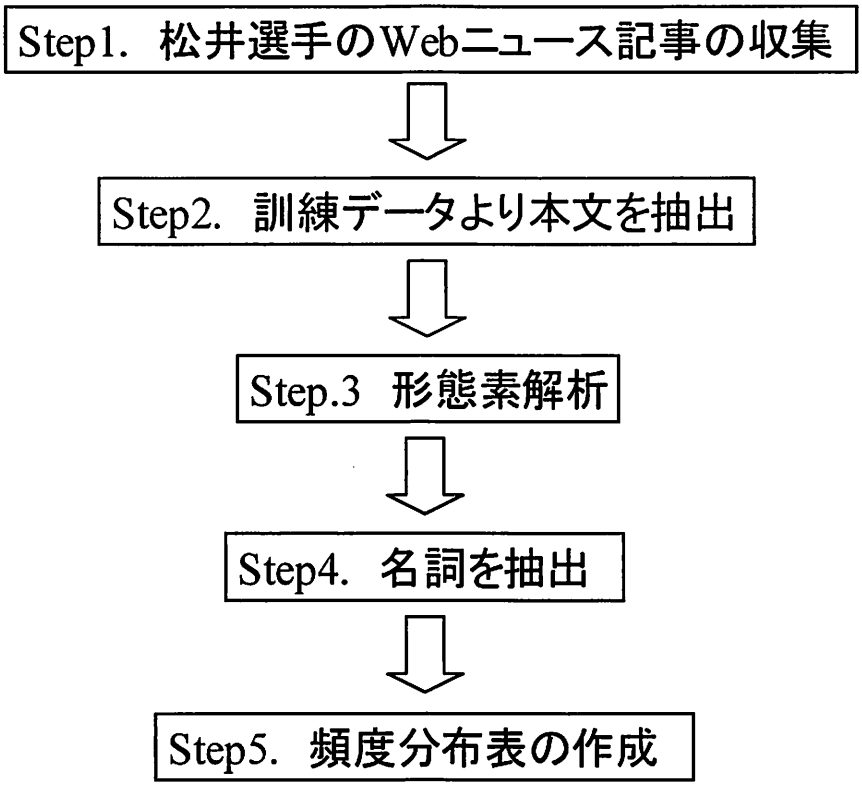

図4.3: 訓練データ作成の流れ.

(2) 松井選手の Web ニュース記事に Naive Bayes 法を用いる。

Step1. (1) で作成した頻度分布を基に、実際に活躍度を算出したい Web ニュース記事 (実験データ) に Naive Bayse 法を用いる。その為に使用する Web ニュース記事から本文をテキストファイルとして抽出する。

Step2. 抽出したテキストファイルを形態素解析する。

Step3. 形態素解析を行ったものから名詞のみを抽出する。Naive Bayes で使用する頻度分布は名詞から作成したためである。

Step4. Naive Bayes 法を適用する。これにより Web ニュース記事がどのクラスに属するのかが求まる。今回の研究でのクラスは「○」、「×」の 2 通りである。したがって Naive Bayes 法の $P(c)$ である $P(○)$ と $P(×)$ は等しくなる。そこで式 (4.3) を算出することにより文章を「○」、「×」の 2 つのクラスに分類することができる。

実験データに Naive Bayes 法を用いる流れを図 4.4 に示す。

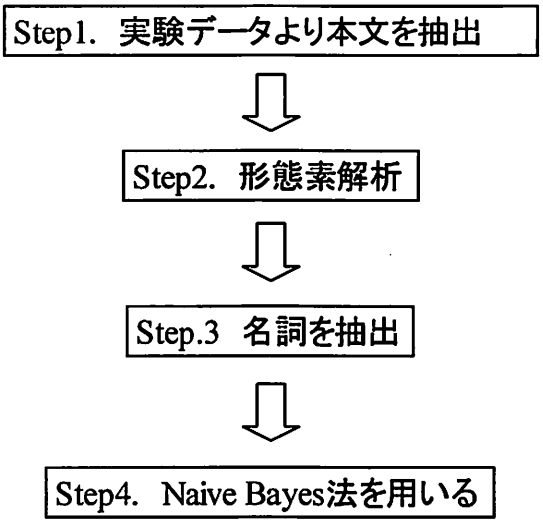

図 4.4: Naive Bayes 法を用いる流れ.

(3) 活躍度を算出する。

Naive Bayes 法を用いて Web ニュース記事におけるを算出する。Naive Bayes 法を用いただけではどのクラスに所属するかしか分からない。したがって Web ニュース記事がどのクラスに近いかを割合を使って表す。これを活躍度とする。

Step1. Naive Bayes 法のところで説明しように、Naive Bayes 法で使用する式 (4.3) の $P(x|c) = \prod_{i=1}^{n} P(f_i|c)$ は文章中の単語の出現の確率による積であるので非常に小さい値になってしまう。そこで、そのままでは扱いにくいので、$P(x|c)$ の対数を用いて扱うことにした。

Step2. Step1. で $\log P(x|c)$ を用いるとしたため、計算結果は負の値となる。このまま比をとると大きさが逆転してしまうために、比の逆数を用いる。$\log P(x|\bigcirc)$ の比の逆数を α とし、$\log P(x|\times)$ の比の逆数を β とする。

$$\alpha = \frac{|\log P(x|\bigcirc)| + |\log P(x|\times)|}{|\log P(x|\bigcirc)|} \tag{4.6}$$

$$\beta = \frac{|\log P(x|\bigcirc)| + |\log P(x|\times)|}{|\log P(x|\times)|} \tag{4.7}$$

Step3. α と β を用いて、クラス「○」かクラス「×」のどちらにどれだけ近いかという割合を用いた活躍度を算出する。訓練データとして使用した Web ニュース記事に Naive Bayes 法を適用する際に、訓練データより単語の頻度分布を作成しているので、どちらのクラスに属するかという推定は明らかなものが現れると考えられる。訓練データに対する α、β の最大値、最小値を求め、実験データに対する α または β を代表して変数 x とおくと、活躍度 y は式 (4.8) で計算される。ただし ε は微小な定数である。

$$x - (\text{最小値} - \varepsilon) : (\text{最大値} + \varepsilon) - (\text{最小値} - \varepsilon) = y : 100 \tag{4.8}$$

この式によって、クラス「○」、「×」どちらにどれだけ近いかという割合を表す活躍度 y を算出する。

Naive Bayes 法を用いて活躍度を算出する流れを図 4.5 に示す。

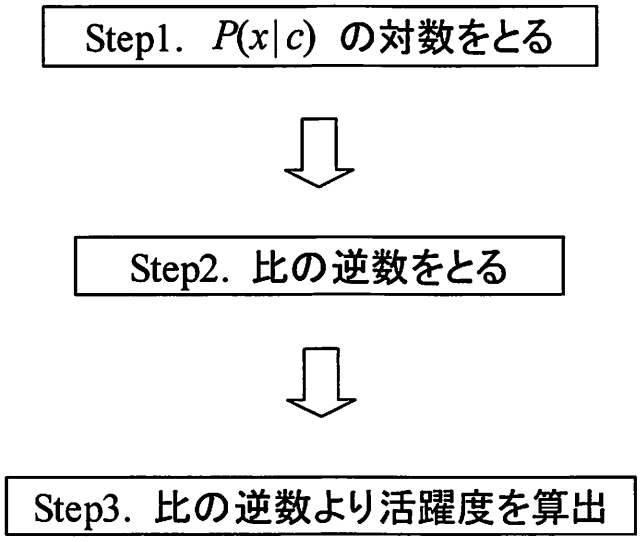

図 4.5: 活躍度算出の流れ.

第5章　実験

松井選手の日本での取り上げられ方はスポニチアネックス(http://www.sponichi.co.jp/)のWebニュース記事を、アメリカでの取り上げられ方はNEW YORK POST(http://www.nypost.com/)のWebニュース記事を使い、日本のWebニュース記事とアメリカのWebニュース記事の内容から活躍度を算出し比較する。

5.1　日本のWebニュース記事より活躍度の算出

5.1.1　訓練データより頻度分布表作成

まず訓練データを収集し、頻度分布表を作成する。日本はスポニチアネックスの2003年のWebニュース記事より松井選手が活躍した(○)、または活躍できなかった(×)と分かる記事を、それぞれ12記事づつ集め、これを訓練データとした。集め方として、特徴があり内容的に似か寄らないように収集することに注意した(表5.1)。幅広くから使用する単語を収集するためである。これはNaive Bayes法の学習は$P(c)$と$P(f_i|c)$の推定であり、訓練データから得られるクラスの頻度や素性の単語とクラスの頻度によって推定されるからである。

表 5.1: 訓練データに選んだ特徴.

日付	「○」に選んだ特徴	日付	「×」に選んだ特徴
4/17	5 打数 3 安打 7 打点、満塁男	5/17	無安打、タイムリーエラー
4/30	5 度目の V 打、超美技	5/24	ゴロ王、バッシング始まる
6/5	ゴジラ大噴火、4 号	5/26	無安打、2 三振、オーナー苦言
6/28	ダブルヘッダー、大暴れ	5/31	無安打、最悪の日
7/4	チーム 2 冠王	6/4	降格検討
7/5	孤軍奮闘	7/10	3 試合連続無安打、どうした松井
7/17	日本人初メジャーサヨナラ弾	8/3	サヨナラ負け演出、牽制死
8/14	捨て身のファインプレー	8/6	外野のリーグ失策王
9/11	2 戦連続マルチ安打、ホームでの捕殺	8/10	打てず、守れず …
9/28	新人 163 試合出場	8/22	併殺打リーグワーストタイ
10/19	オクトーバー・モンスター	8/26	先発落ち
10/21	2 試合連続勝利打点	10/15	松井のミスで … 逆転負け

これらの特徴より選んだ 24 個の Web ニュース記事より本文をテキストファイルとして抽出し、形態素解析を行い名詞のみを取り出した。そして、活躍できたクラス「○」と活躍できなかったクラス「×」という 2 つのクラスでそれぞれ何回その名詞が使用されたかという頻度分布表を作成した (表 5.2)。

表 5.2: 訓練データの頻度分布表.

名詞	「○」の数	「×」の数
本塁打	12	10
凡退	0	3
毎日	0	1
幕	3	1
…	.	.
…	.	.
…	.	.

5.1.2 活躍度の算出

まず訓練データとして収集した24個のWebニュース記事にNaive Bayes法を適用する(表5.3)。今回の研究で使用するNaive Bayes法は活躍できたというクラス「○」と、活躍できなかったというクラス「×」の2つのクラスに分類する。そして、クラス「○」か「×」のどちらにどれだけより近いか分かる為に比の逆数を算出した。

表5.3: 訓練データの値.

○ or ×	日付	logP(x—○)	logP(x—×)	比の逆数「○」	比の逆数「×」
○	4/17	-1680.31462	-2176.24924	2.295144	1.772115
○	4/30	-1852.94955	-2392.99167	2.29145	1.774323
×	5/17	-1157.32084	-886.405057	1.765911	2.305634
×	5/24	-793.953986	-607.003572	1.764532	2.307989
×	5/26	-2742.85975	-2046.45812	1.746104	2.340296
×	5/31	-831.875834	-665.456773	1.799947	2.250082
×	6/4	-1625.5201	-1246.05444	1.766557	2.304534
○	6/5	-1672.03157	-2051.39746	2.226889	1.81507
○	6/28	-1537.78486	-1948.59177	2.267142	1.789176
○	7/4	-1491.92656	-1992.9946	2.335853	1.748585
○	7/5	-1651.97075	-2133.4156	2.291437	1.774331
×	7/10	-1435.61309	-1062.593	1.740167	2.351047
○	7/17	-1453.70945	-1846.70298	2.270338	1.787192
×	8/3	-1608.1088	-1281.65337	1.796994	2.254714
×	8/6	-1750.53546	-1361.74569	1.777902	2.285508
×	8/10	-1492.18984	-1191.19019	1.798283	2.252688
○	8/14	-1184.15076	-1541.4403	2.301726	1.768211
×	8/22	-1023.52612	-857.723783	1.838009	2.193305
×	8/26	-1473.71196	-1211.14811	1.821835	2.216789
○	9/11	-1247.2021	-1650.81648	2.323616	1.755506
○	9/28	-1225.41316	-1541.85068	2.258229	1.794768
×	10/15	-2082.77732	-1658.97621	1.796521	2.255459
○	10/19	-1893.9293	-2560.58347	2.351995	1.739648
○	10/21	-1213.41716	-1537.17214	2.266813	1.789383

さらに「○」か「×」の割合をパーセントで表した活躍度を算出する。式(4.6)と式(4.7)を用いて訓練データより求めたαとβの値の中での最大値2.351995(表5.3の10/19日)と、最小値1.739648(表5.3の10/19日)を式(4.8)に代入する。微小な定数εは$\varepsilon = 0.004$として活躍度yを算出する(表5.4)。

$$(x - 1.735648) : (2.355995 - 1.735648) = y : 100$$

$$y = \frac{100(x - 1.735648)}{0.620347} \tag{5.1}$$

表 5.4: 訓練データの活躍度.

○ or ×	日付	活躍度「○」	活躍度「×」
○	4/19	90.190797	5.87848
○	5/2	89.595338	6.234485
×	5/19	4.8784401	91.88185
×	5/26	4.6561734	92.26143
×	5/28	1.6854781	97.46933
×	6/2	10.365043	82.92689
×	6/6	4.9825963	91.70444
○	6/7	79.18813	12.80276
○	6/30	85.676885	8.628968
○	7/6	96.753111	2.085502
○	7/7	89.59319	6.235773
×	7/12	0.7284144	99.20238
○	7/19	86.192152	8.308867
×	8/5	9.8890116	83.67354
×	8/8	6.8114099	88.63754
×	8/12	10.096813	83.34693
○	8/16	91.251895	5.249093
×	8/24	16.500554	73.77436
×	8/28	13.893358	77.56001
○	9/13	94.78048	3.201145
○	9/30	84.240151	9.530106
×	10/17	9.8127598	83.79366
○	10/21	99.355244	0.644729
○	10/23	85.623786	8.662046

2003年8月15日から10月20日(現地の試合の日にち)のWebニュース記事にNaive Bayes法を用いた(表5.5)。ただし9月16日と9月28日は記事が更新されなかったかめ、活躍度は0とした。

表5.5: 実験データの値.

日付	$logP(x\|\bigcirc)$	$logP(x\|\times)$	比の逆数「○」	比の逆数「×」
8/15	-685.113239	-734.340839	2.071853231	1.932963554
8/16	-428.766655	-461.991213	2.07748867	1.928084004
8/17	-751.386387	-804.615129	2.070840706	1.933845711
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
9/16	0	0	0	0
9/17	-898.959922	-919.923831	2.023320182	1.977211256
9/18	-855.814904	-963.855696	2.126243177	1.88790771
9/19	-838.153928	-861.170441	2.027460962	1.973272988
9/22	-418.01994	-387.424666	1.926809056	2.078970899
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
10/18	-478.00265	-508.94095	2.064724118	1.939210433
10/19	-1706.81248	-1735.494	2.016804142	1.98347357
10/20	-1436.83978	-1563.33533	2.088037338	1.919086106

訓練データでの活躍度のように実験データの活躍度は、クラス「○」の活躍度とクラス「×」活躍度の差がはっきりと表れなかった。これは訓練データよりもクラス「○」とクラス「×」の比の逆数の差が小さいため、活躍度に差がでなかった。そこで、8月15日から10月20日のクラス「○」とクラス「×」の逆数の比の中で、式(4.8)と同様に最大値2.126243177(表5.5より9月18日)、最小値1.88790771(表5.5より9月18日)を用いて新たに活躍度を求めた(表5.6)。

$$x - (\text{最小値} - \varepsilon) : (\text{最大値} + \varepsilon) - (\text{最小値} - \varepsilon) = z : 100$$

$$(x - 1.885) : 0.243 = z : 100$$

$$z = \frac{100(x - 1.885)}{0.243} \tag{5.2}$$

表5.6: 実験データの活躍度.

日付	活躍度「○」	活躍度「×」	新活躍度「○」	新活躍度「×」
8/15	54.19631775	31.80728753	76.89433	19.73809
8/16	55.10475111	31.02070365	79.21344	17.73004
8/17	54.03309854	31.94949129	76.47766	20.10112
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
9/16	0	0	0	0
9/17	46.37278523	38.94002158	56.92189	37.94702
9/18	62.9639825	24.54428089	99.27703	1.196589
9/19	47.04027936	38.30517237	58.62591	36.32633
9/20	52.4582932	33.33547262	72.45739	23.63934
9/21	49.88840212	35.65286782	65.89679	29.55535
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
10/18	53.04710399	32.81428506	73.96054	22.30882
10/19	45.32239899	39.94950729	54.24039	40.5241
10/20	56.80519734	29.57024153	83.55446	14.0272

5.2 アメリカのWebニュース記事より活躍度の算出

アメリカのWebニュース記事については松井選手の名前が挙がった日に活躍度があると考えた(表5.7)。これは大活躍できた、または全く活躍できなかったというように、何かしら注目された特徴的な日であったと考えたからである。

表5.7: アメリカサイトにおける松井選手の名前が挙がった日.

日付	日付	日付
8/17	9/6	10/1
19	7	2
22	8	6
26	22	7
29	25	8
31	29	9
-	30	10
-	-	11
-	-	12
-	-	16
-	-	17
-	-	18

5.3 日本とアメリカの活躍度の比較

アメリカのWebニュース記事で名前が挙がった日と、日本のWebニュース記事のNaive Bayes法を用いて表した割合の活躍度とを比べることにより日本とアメリカの活躍度の比較を行う方法をとることにした。日本サイトにおいける活躍度とアメリカサイトに名前が挙がった日との比較結果を図5.1に示す。ただし、9月16日と9月28日は日本の活躍度が0となっているがこれは記事がなかった為である。

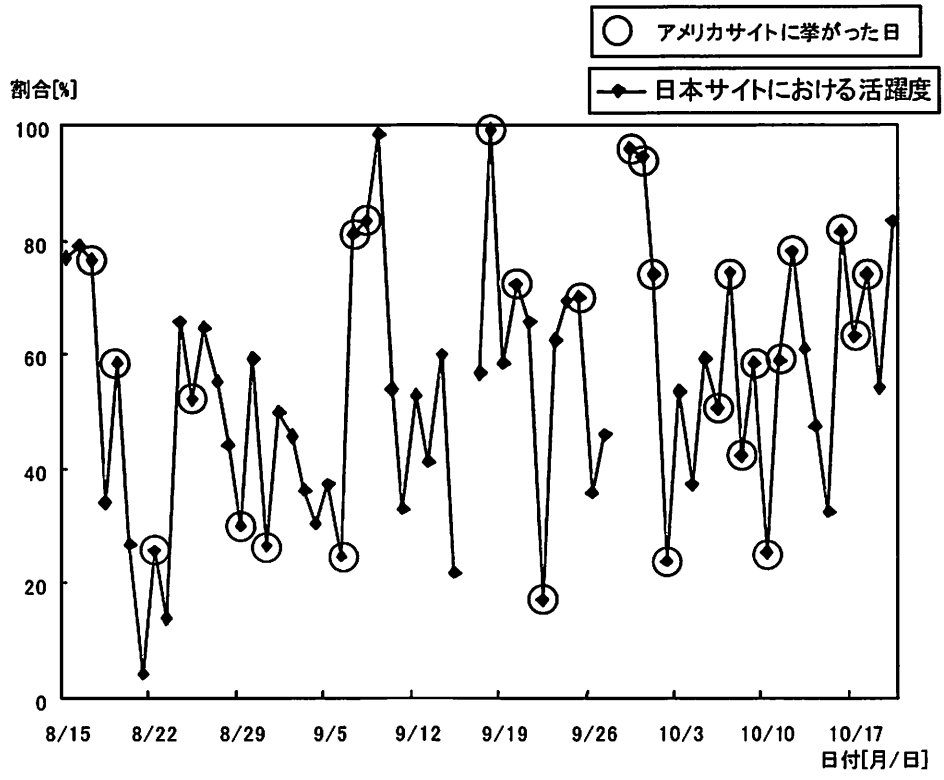

図 5.1: 実験結果の比較 1.

第6章　考察

6.1　Naive Bayes法による結果

まず、Naive Bayes法を用いて実験データを「○」、「×」の二つのクラスに別けた場合について考える。実際に記事を読み松井選手が活躍できたまた、活躍できなかったという○、×を付けたものとNaive Bayes法による分類との結果を比べると、約85%の正答率が約85%であることが分かった。これは、Webニュースの記事のクラス「○」かクラス「×」であるかの分類においては既存の研究と同程度の結果である。

次に、式(5.2)による活躍度ついて考える。実験データより活躍度の値が妥当な数であるかという正答率において、記事を読み検証を行った。特に活躍度が高い日か低い日、図6.1ので70%以上の日と30%未満になっている日の活躍度を見ると、正答率は約75%になった。活躍度が高いと低い日の正答率が高かった。しかし1つ1つをチェックすると活躍度30%以上70%未満の活躍度の妥当性であるかの正答率は約35%である。このような結果となったのは、頻度分布表に大きな原因がある。Naive Bayes法を利用して活躍度は算出されているため、クラス「○」とクラス「×」で使用されている名詞の出現回数に差が大きいものが多く含まれている記事は活躍度が高いまたは低いものとなり、またそうでないものが活躍度30%以上70%未満に集まる結果となったためである。

6.2　アメリカの活躍度の結果

アメリカのWebニュース記事の中に松井選手の名前が挙がった日は大活躍した、または全く活躍できなかったという特徴的な日であると考えた。実際に松井選手の名前が挙がった日の記事を読むと松井選手が活躍したまたは、できなかったものが多かった。しかし、その中には名前が挙がっているだけで活躍を評価していない記事もみられた図6.1。活躍したまたは活躍できなかったという記事の内容に

おいては日本サイトの記事と類似していた。しかし、日本のサイトほど情報は豊富ではなかった。これは使用したNEW YORK POST(http://www.nypost.com/)の記事が松井選手中心に書かれた記事ではなくチーム全体について書かれているためである。そのような中、松井選手が大活躍した、プレーオフにおいては毎日のように名前が挙がり、日本と同様に記事がたくさん載っていた。このように、松井選手が大活躍または全く活躍できなかった日のアメリカのWebニュース記事には松井選手の名前が挙がっていたことより、名前が挙がった日は大活躍した、または全く活躍できなかったと考えたことは、良い結果である思われる。

アメリカのWebニュース記事については日本のWebニュース記事と同様の方法を使って活躍度を算出することができなかった。使用したNEW YORK POSTの記事が日本のサイトの書き方とは違い、松井選手中心書かれたものではなかった為、訓練データのクラスの分類ができないのがその理由である。また、日本サイトは2003年の記事を訓練データとして使用したが、アメリカサイトは過去のデータを収集することができず、訓練データの収集ができなかったのも理由である。さらに、品詞ごとに分割する形態素解析を行うことができなかった為、実験としてNaive Bayes法を適用することができなかった。

6.3 日本とアメリカの活躍度の比較

日本とアメリカのWebニュース記事の活躍度の比較を行う。図6.1を見る限りアメリカサイトで名前が挙がり評価が高かった日と低かった日は、日本サイトの活躍度は高いか低いかに別れていることがわかる。これは、日本のWebニュース記事では大活躍したのと、全く活躍できなかったという両極端の記事については、高い正答率となり、アメリカサイトにおいても両極端の日に記事に挙がっていたためにこのような結果となった。しかし、Webニュース記事の内容を活躍度として正しく算出されていないので、アメリカでは低い評価であった記事が日本では活躍度が高いものやアメリカにおいては名前が挙がっただけであるのに対し、日本では活躍度が高いという結果の日が見られた。

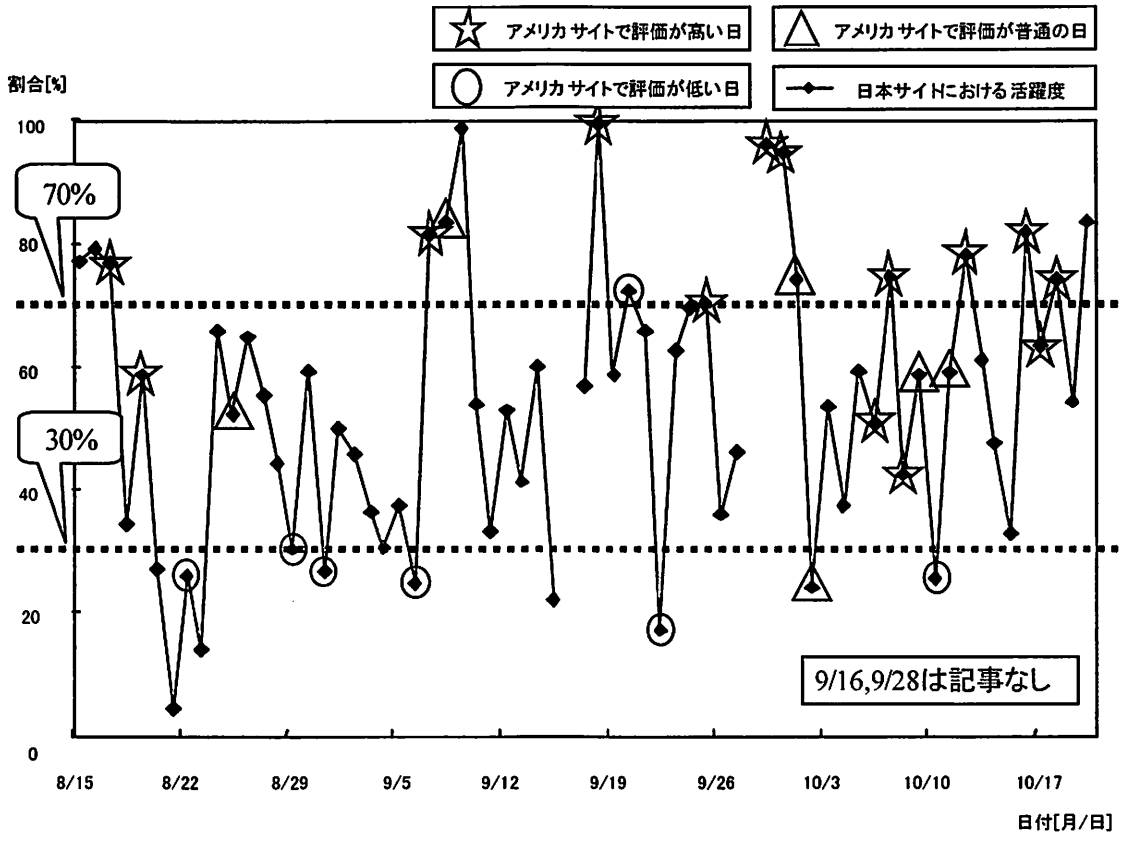

図 6.1: 実験結果の比較 2.

第7章　まとめ

今回の実験ではNaive Bayes法を用いて日本とアメリカのWebニュース記事より活躍度を算出して比較するという実験を行ったが、これは日本サイトのみしか行うことができなかった。これはアメリカサイトにおいても行うことができれば良かった。しかし、日本とアメリカの評判は定性的には類似していた。もしアメリカのWebニュース記事においてNaive Bayes法を適用していたら日本での結果のように、「○」か「×」という分類においては良い結果を残すだろう。しかし、活躍度として算出するためにはさらに精度向上を測らなくてはならない。そこで、訓練データより単語を抽出し頻度分布表を作成する際、名詞のみではなく評価を表す形容詞や、形容動詞を用いることにより精度向上を測るということが考えられる。また訓練データを増やすことも有効であると考えられる。最後に今回の実験で日本での松井選手に対する評判とアメリカでの評判では、日本の方が全体的に評価が高いがアメリカにおいても高い評価であった。

第8章　謝辞

本研究の遂行及び論文の作成に多大な御助言及び指導を賜わった新納浩幸教官 (茨城大学工学部システム工学科) に深い感謝の意を表します。

最期に、本研究を進めるにあたり助言、強力頂きました、岩崎唯史教官 (茨城大学工学部システム工学科)、同研究室の紺野憲一氏 (茨城大学大学院修士課程)、藤井丈明氏 (茨城大学大学院修士課程)、正木裕一氏 (茨城大学大学院修士課程)、谷津哲平氏 (茨城大学大学院修士課程)、木本俊氏 (茨城大学工学部システムシステム工学科４回生)、藤村元彦氏 (茨城大学工学部システムシステム工学科４回生)、茂木啓悟氏 (茨城大学工学部システムシステム工学科４回生)、及び、岩崎研究室の高橋宏直氏 (茨城大学工学部システムシステム工学科４回生)、にも深く感謝します。

付録A　実験データ

表 A.1: 実験データ 1

日付	$logP(x\|○)$	$logP(x\|×)$	逆数の比「○」	逆数の比「×」
8/15	-685.113239	-734.340839	2.071853231	1.932963554
8/16	-428.766655	-461.991213	2.07748867	1.928084004
8/17	-751.386387	-804.615129	2.070840706	1.933845711
8/18	-563.346325	-545.172431	1.967739394	2.033336047
8/19	-1065.59942	-1095.26615	2.027840412	1.972913682
8/20	-540.770819	-513.967495	1.950434966	2.052149843
8/21	-553.579207	-495.794511	1.895616209	2.116549689
8/22	-1199.35956	-1136.65541	1.947718642	2.055165484
8/23	-398.057387	-365.944727	1.919326557	2.087752761
8/24	-618.710486	-646.531225	2.044965682	1.956969226
8/25	-513.793473	-519.73207	2.011558335	1.988573734
8/26	-866.090542	-902.708057	2.042279084	1.959435928
8/27	-1057.25889	-1077.99531	2.019613377	1.980763908
8/28	-934.610555	-927.957019	1.992880954	2.007170091
8/29	-657.427959	-629.916626	1.958153083	2.043674562
8/30	-368.97837	-379.718833	2.029108652	1.971714695
8/31	-610.232173	-579.486213	1.949615964	2.053057276
9/1	-477.188029	-480.472122	2.006882178	1.993164863
9/2	-388.400728	-386.733207	1.9957067	2.004311812
9/3	-750.695024	-730.819456	1.973523778	2.027196277
9/4	-586.973722	-563.002707	1.95916169	2.042577087

9/5	-634.501297	-619.173041	1.975842042	2.024756013
9/6	-915.847891	-865.888772	1.945450419	2.057696924
9/7	-677.13279	-732.858559	2.082296663	1.923961086
9/8	-571.209722	-621.58825	2.088196202	1.91895193
9/9	-1102.94976	-1240.67661	2.124871371	1.888990534
9/10	-903.80755	-918.516515	2.016274444	1.983986173
9/11	-481.911685	-465.389914	1.965716185	2.035500922
9/12	-1034.59845	-1048.25662	2.013201423	1.986970584
9/13	-730.121945	-719.846667	1.985926628	2.014274259
9/14	-505.103042	-520.839226	2.031154404	1.969786869
9/15	-847.91495	-795.41251	1.938080535	2.066006556
9/16	0	0	0	0
9/17	-898.959922	-919.923831	2.023320182	1.977211256
9/18	-855.814904	-963.855696	2.126243177	1.88790771
9/19	-838.153928	-861.170441	2.027460962	1.973272988
9/20	-516.007725	-547.521064	2.061071448	1.942443604
9/21	-573.209665	-599.078162	2.045129206	1.956819496
9/22	-418.01994	-387.424666	1.926809056	2.078970899
9/23	-873.340873	-906.050672	2.037453645	1.963898488
9/24	-1335.46616	-1407.46487	2.053912794	1.948845109
9/25	-1057.71662	-1116.23448	2.055324707	1.947575655
9/26	-434.819469	-422.749405	1.972241206	2.028551345
9/27	-504.615152	-502.832344	1.996466995	2.003545532
9/28	0	0	0	0
9/29	-1115.15839	-1247.55563	2.118725061	1.893874675
9/30	-1303.20075	-1453.32756	2.11519853	1.896701326
10/1	-583.980825	-621.785531	2.064736211	1.939199766
10/2	-518.911311	-489.835154	1.943967001	2.059359066

10/3	-1049.17779	-1065.07495	2.015152014	1.985074142
10/4	-1536.04713	-1499.30058	1.976077203	2.024509124
10/5	-964.774553	-993.345712	2.029614337	1.971237447
10/6	-1398.64918	-1410.62561	2.008562848	1.991509851
10/7	-761.569442	-811.721541	2.065853613	1.938215143
10/8	-1211.30882	-1196.64391	1.987893338	2.01225503
10/9	-1559.66868	-1602.23099	2.027289326	1.973435599
10/10	-435.93405	-412.6707	1.946635621	2.056372672
10/11	-983.019087	-1011.13907	2.028605736	1.972189796
10/12	-1314.76015	-1413.07484	2.074777662	1.930424994
10/13	-979.162684	-1012.02044	2.033556998	1.967532514
10/14	-589.290249	-589.382968	2.00015734	1.999842685
10/15	-682.725728	-658.544948	1.964582	2.036718496
10/16	-1309.78455	-1418.4772	2.082985139	1.923373705
10/17	-1130.43034	-1174.44179	2.038933359	1.962525644
10/18	-478.00265	-508.94095	2.064724118	1.939210433
10/19	-1706.81248	-1735.494	2.016804142	1.98347357
10/20	-1436.83978	-1563.33533	2.088037338	1.919086106

表 A.2: 実験データ 2

日付	活躍度「○」	活躍度「×」	新活躍度「○」	新活躍度「×」
8/15	54.19631775	31.80728753	76.89433	19.73809
8/16	55.10475111	31.02070365	79.21344	17.73004
8/17	54.03309854	31.94949129	76.47766	20.10112
8/18	37.41315646	47.98734376	34.04913	61.04364
8/19	47.10144678	38.24725229	58.78206	36.17847

8/20	34.62368095	51.02012949	26.92797	68.78594
8/21	25.78689176	61.4013913	4.36881	95.28794
8/22	34.1858093	51.50625118	25.81014	70.02695
8/23	29.60900219	56.7593235	14.12616	83.43735
8/24	49.86204206	35.67700433	65.8295	29.61697
8/25	44.47677426	40.77165421	52.08162	42.62294
8/26	49.42896222	36.07463687	64.7239	30.63207
8/27	45.77524786	39.51270948	55.39645	39.40902
8/28	41.46597848	43.76938886	44.39545	50.27576
8/29	35.86784225	49.65391344	30.10415	65.29817
8/30	47.3058872	38.05397544	59.30397	35.68506
8/31	34.49165767	51.16640785	26.59093	69.15937
9/1	43.72297732	41.51174464	50.15727	44.51229
9/2	41.92148904	43.30863409	45.55831	49.09951
9/3	38.34559982	46.99761206	36.42954	58.51699
9/4	36.03042969	49.47700022	30.51921	64.84654
9/5	38.71930412	46.60424137	37.38356	57.51276
9/6	33.8201714	51.91431963	24.87672	71.06869
9/7	55.87980001	30.35608876	81.19204	16.03337
9/8	56.8308063	29.54861231	83.61984	13.97199
9/9	62.74284727	24.71883212	98.7125	1.642195
9/10	45.23701148	40.03213887	54.0224	40.73505
9/11	37.08701503	48.33632176	33.21654	61.93454
9/12	44.74164028	40.51322624	52.75779	41.9632
9/13	40.34494043	44.91458157	41.53359	53.19928
9/14	47.63566255	37.74320961	60.14585	34.89172
9/15	32.63214541	53.25383302	21.84384	74.48829
9/16	0	0	0	0

9/17	46.37278523	38.94002158	56.92189	37.94702
9/18	62.9639825	24.54428089	99.27703	1.196589
9/19	47.04027936	38.30517237	58.62591	36.32633
9/20	52.4582932	33.33547262	72.45739	23.63934
9/21	49.88840212	35.65286782	65.89679	29.55535
9/22	30.81518183	55.34368645	17.20537	79.82342
9/23	48.65110091	36.7940021	62.73813	32.46851
9/24	51.30431746	34.36739581	69.51144	26.27371
9/25	51.53191793	34.16275965	70.09247	25.7513
9/26	38.13884909	47.2160492	35.90173	59.07463
9/27	42.04404869	43.18510957	45.87119	48.78417
9/28	0	0	0	0
9/29	61.75206145	25.50615623	96.18315	3.65213
9/30	61.18358431	25.96181261	94.73191	4.81536
10/1	53.04905338	32.81256549	73.96552	22.30443
10/2	33.58104431	52.18225695	24.26626	71.7527
10/3	45.05607578	40.20751972	53.5605	41.18277
10/4	38.75721213	46.56444282	37.48033	57.41116
10/5	47.38740371	37.97704296	59.51207	35.48866
10/6	43.99390155	41.24495668	50.84891	43.83121
10/7	53.2291786	32.65384429	74.42535	21.89924
10/8	40.66197427	44.58908159	42.34294	52.36833
10/9	47.01261164	38.33138529	58.55528	36.39325
10/10	34.01122609	51.70085	25.36445	70.52373
10/11	47.224817	38.13056174	59.09701	35.88057
10/12	54.6677363	31.39807143	78.0978	18.69341
10/13	48.02296098	37.37980737	61.13457	33.964
10/14	42.63893275	42.58821025	47.38985	47.26036

10/15	36.90418423	48.53259481	32.74979	62.4356
10/16	55.99078246	30.26140281	81.47537	15.79165
10/17	48.88963096	36.57269941	63.34706	31.90356
10/18	53.04710399	32.81428506	73.96054	22.30882
10/19	45.32239899	39.94950729	54.24039	40.5241
10/20	56.80519734	29.57024153	83.55446	14.0272

付録B　プログラムソースリスト

```
%%%%%%%%%%%%%%%%%%%%%%%%%%
本文抽出のためのプログラム
%%%%%%%%%%%%%%%%%%%%%%%%%%
#include<stdio.h>
#include<string.h>
#include<stdlib.h>

#define LINESIZE 15000
#define WORDSIZE 11000
void quit(char *);

int main(int argc, char *argv[])
{
  FILE *f1;
  char buf[LINESIZE],str[LINESIZE],title[WORDSIZE];
  char *key1 = "<!--ニュース見出しここから -->";
  char *key2 = "<!--P説ここから --> ";
  char *key3 = "<!--P説ここまで -->";
  char *key4 = "ニュースの詳細はスポーツニッポン新聞本紙";
  char *cp1,*cp2,*cp3,*cp4;
  int i,j,k;
  int flag1=0,flag2=0,flag0=0,flag3=0;

  if(argc != 2)quit("引数が違う prog datafile");
```

```
  if((f1 = fopen(argv[1],"r")) == NULL)quit("ファイルが開けな
い");
  strcpy(buf,"");
  strcpy(str,"");
  strcpy(title,"");
  i=j=k=0;

  while(fgets(buf,LINESIZE,f1) != NULL)
  {
    i=j=0;
    flag3=0;


      if(strstr(buf,"速報一覧")!=NULL)
      {
         flag1=flag2=0;
      }
         if((cp2 = strstr(buf,key2)) != NULL) flag1=0;
         if((cp3 = strstr(buf,key3)) != NULL) flag2=1;
         if((cp4 = strstr(buf,key4)) != NULL) flag2=0;
         if((cp1 = strstr(buf,key1)) != NULL) flag1=1;
         if(strstr(buf,"  ")!=NULL) flag3=1;

         if(flag1==1 || flag2==1)
 {
             while(buf[i] != '\0')
             {

                if(buf[i] == '<')
       {
          flag0=1;
```

```
        }

                if(flag0 == 0)
                {
                   if(isspace(buf[i])==0)
                   {
              title[j]=buf[i];  /* <,>のどちらに入れても同じ */
                     j++;
                   }
                }
                if(buf[i] == '>')
        {
                  flag0=0;
                }

                i++;
      }
title[j]='\0';   /* NULL文字を入れる */
        if(strcmp(title,"\0")!=0 )printf("%s\n",title);
strcpy(title,"");
        } /* flag1==1,flag2==1 */
  }  /* while(一行読み込み)のかっこ閉じ */
}  /* mainのかっこ閉じ */


void quit(char *s)
{
   puts(s);exit(1);
 }
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%
名詞抽出のためのプログラム
%%%%%%%%%%%%%%%%%%%%%%%%%
#include<stdio.h>
#include<stdlib.h>
#include<string.h>

#define LINESIZE 10000
#define WORDSIZE 100

void quit(char *);

int main(int argc, char *argv[])
{
  FILE *f1;
  char buf[LINESIZE],word1[WORDSIZE],word2[WORDSIZE],word3[WORDSIZE],
  int r;

  if(argc != 2) quit("引数の数が違う prog datafile");
  if((f1 = fopen(argv[1],"r")) == NULL) quit("ファイルが開けな
い");
  while(fgets(buf,LINESIZE,f1) != NULL)
    {
      sscanf(buf,"%s %s %s %s",word1,word2,word3,word4);
      if(strcmp(word4,"名詞") == 0)
        {
  if(strcmp(word1,"EOS") != 0)
          puts(word1);
}
    }
```

```
  if((r = fclose(f1)) == -1) quit("ファイルが閉じれない");
}

void quit(char *s)
{
  puts(s); exit(1);
}
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%
Naive Bayes法のプログラム
%%%%%%%%%%%%%%%%%%%%%%%%%
#include<stdio.h>
#include<stdlib.h>
#include<string.h>
#include<math.h>

#define LINESIZE 256
#define WORDSIZE 100

void quit(char *);

int main(int argc, char *argv[])
{
  FILE *f1,*f2;
  char str[LINESIZE],buf[LINESIZE],word1[WORDSIZE],word2[WORDSIZE],wc
  int r;
  double wari1=0,wari2=0,total1=0,total2=0,n1,n2,s1=2928,s2=2254,hi1,

  if(argc != 2) quit(" 引数の数が違う prog datafile ");
```

```
  if((f1 = fopen(argv[1],"r")) == NULL ) quit("ファイルが開けない");


  while(fgets(buf,LINESIZE,f1) != NULL)    // ファイルより一行読み込み
    {
      f2 = fopen("tekitou.txt","r"); // 頻度分布表のファイル名
      while(fgets(str,LINESIZE,f2) != NULL)
{
          sscanf(str,"%s %s %s",word1,word2,word3);    // データより一行読み込み
  strcat(word1,"\n");
  if(strcmp(buf,word1) == 0)
    {
      n1 = atoi(word2);
      n2 = atoi(word3);

      wari1 = n1 / s1;
      wari2 = n2 / s2;

            if(wari1 == 0)
        {
          wari1 = 0.000001;
        }

            if(wari2 == 0)
        {
          wari2 = 0.000001;
        }
```

```
                total1 = total1 + log(wari1);
                total2 = total2 + log(wari2);
                wari1 = wari2 = 0;
    }
}
      fclose(f2);



    }
  printf("対数は %f %f\n",total1,total2);

  if((r = fclose(f1)) == -1) quit("ファイルが閉じれない");
}

void quit(char *s)
{
  puts(s); exit(1);
}
```

参考文献

[1] 藤村滋, 豊田正史, 喜連川優 : “Web からの評判および評価表現抽出に関する一考察”, 電子情報通信学会データ工学研究専門委員会, 2004

[2] 垣東伸明: “評価表現を用いた商品評価ページのマッピング” 平成 15 年度卒業研究, 2004

[3] 藤村滋, 豊田正史, 喜連川優: “電子掲示板からの評判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, 2004.6

[4] 阿部修也: “帰納学習における背景知識の利用”, 平成 14 年度修士学位論文, 2002