社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－NL－173 (12)
2006／5／19

野球チーム名における固有名詞の反復出現について

高瀬 暁央† 梅村 恭司‡

† ‡豊橋技術科学大学 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
E-mail: †takase@ss.ics.tut.ac.jp, ‡umemura@tutics.tut.ac.jp

抄録 キーワードの分布として良く知られているものに、Katz K mixture モデルがある。この Katz K mixture モデルは、キーワードが文書中で繰り返し出現する条件付き確率は減少係数によって決められる、と仮定している。しかし、この Katz K mixture モデルに合致しないキーワードが存在する。その一つが日本のプロ野球で使われているチーム名である。野球チーム名には地名や企業名などが含まれているが、野球チーム名として使われていない地名や企業名を調べた結果、野球チーム名だけが特異な特徴を持つことがわかった。本研究では、新聞記事中に出現する野球チーム名が Katz K mixture モデルと合致せず、また特異な特徴を持っているという発見を報告する。

キーワード Katz モデル 統計的言語処理 単語頻度 固有名詞

How Repeatedly Baseball Team Names appear in an Article

Akihiro TAKASE† Kyoji UMEMURA‡

† ‡ Toyohashi University of Technology 1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, Aichi, 441-8580 Japan
E-mail: †takase@ss.ics.tut.ac.jp, ‡umemura@tutics.tut.ac.jp

Abstract The Katz K Mixture model is well known model for keywords and proper nouns. This model assumes that there are constant decay factors for the conditional probability of repeats. We have found that there are some keywords that do not obey this assumption. They are the names of professional baseball teams. We have checked that other names, such as companies or places which obey Katz model, and we have found that the names of baseball teams alone have this special feature. This paper reports the detailed distribution of these baseball names, comparing with other names, and showing the difference among them.

Keyword The Katz K Mixture model, statistical natural language processing, term frequency, proper noun

1. はじめに

キーワードの分布モデルを作成することで、様々な応用が考えられる。例えば、キーワードに対する異常出現の判別や、あまり使われなくなった語やだんだんと使われてくる新語などの検出が期待できる。

この分布モデルについて一般的に知られているものに、Katz K mixture モデル[1]がある。Katz K mixture モデルでは、繰り返しの条件付き確率について減衰係数による法則性を述べている。また、Katz K mixture モデルの改良についての報告[2]もされている。

しかし、この Katz K mixture モデルとは違った性質を持つキーワードの一群がある。それは、新聞記事における野球チーム名である。日本のプロ野球チームは 12 球団だが、その全てに同様の特徴が見られた。この特徴は、Katz K mixture モデルや、改良した Katz K mixture モデルとも違う特徴である。

本稿ではこの特徴について報告し、野球チーム名にのみ現れていることを報告する。

2. 記号の定義

以下に、本論文で使用する記号の定義について述べる。

cf(w) : キーワード w のコーパス中での出現数

df(w) : キーワード w が出現した文書数

df(k;w) : キーワード w が k 回以上出現し

た文書数

N：コーパスサイズ

tf(w,d)：文書dにおけるキーワードwの出現数

$cdf(k;w) \equiv \sum_{i \geq k} df(i;w)$：キーワードwに対するdf(i;w)を累積したもの

$Pw(k+1|k) \equiv P(tf(w,X) \geqq k+1 | tf(w,X) \geqq k)$：Katz K mixtureモデルで使われる、繰り返しの条件付き確率。Xは文書に対する確率変数。

3. Katz K mixture モデル

3.1. 概要

Katz K mixture モデルは繰り返しの条件付き確率について、「以前に何回発生しているかからは独立している」と仮定している。

3.2. 定義式

Katz K mixture モデルは、「文書中にキーワードwがちょうどk回発生する確率」として、以下の式によってモデル化される。

$$P_{Katz} = (tf(w,X) = k) = (1-\alpha)\delta_{k,0} + \frac{\alpha}{\beta+1}\left(\frac{\beta}{\beta+1}\right)^k$$

$$\text{where: } \delta_{k,0} = \begin{cases} 1 & \text{iff} \quad k = 0 \\ 0 & \text{otherwise} \end{cases}$$

このうち、αとβは平均λやIDFから以下のように算出される。

$$\text{observed mean: } \hat{\lambda} = \frac{cf}{N}$$

$$\hat{IDF} = \log_2 \frac{N}{df}$$

$$\hat{\alpha} = \frac{\hat{\lambda}}{\hat{\beta}} = \frac{cf}{N} \times \frac{df}{cf - df} = \frac{df}{cdf(2)}$$

$$\hat{\beta} = \hat{\lambda} \times 2^{\hat{IDF}} - 1 = \frac{cf - df}{df} = \frac{cdf(2)}{df}$$

ここで、Katz K mixture モデルの減衰係数 $P_{Katz}(k+1)/P_{Katz}(k)$ は $cdf(2)/cf$ によって決定される。そして、この減衰係数が繰り返しの条件付き確率 $Pw(k+1|k)$ となる。

3.3. Katz K mixture モデル例

Katz K mixture モデルを示す繰り返しの条件付き確率のグラフを図 1に示す。このグラフでは、2 つのキーワードについて実値と Katz K mixture モデルによる推定からの条件付き確率をプロットしている。このグラフからもわかるとおり、Katz K mixture モデルの推定はx軸と平行な直線になる。

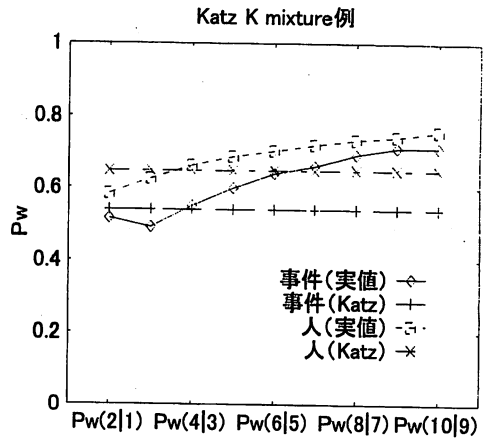

図 1　Katz K mixture モデル例

4. 使用コーパス

本研究で用いたコーパスは、毎日新聞の91年から97年までの計7年、2550日分（毎年1月2は休刊）の、全731548記事を集めた日本語コーパス[3]である。記事の内容は、政治、経済、スポーツ、社会、対談など様々な分野の記事が含まれている。

このコーパスに対して、以下の作業を行った。

1. 茶筌[4]を用いて形態素解析を実行（辞書にはipadic2.7.0を用いた）
2. 記号類を削除
3. 形態素解析によって出力されたキーワードごとのdfをカウント[5]
4. それぞれのキーワードに対して繰り返しの条件付き確率Pw(k+1|k)を算出

この作業の結果、この毎日新聞のコーパスは283088個のキーワードを持ち、のべ185919395回現れている。また、7年分で合計731548件の記事を持ち、記事一つを一文書とした。一つ一つの記事は1個から4455個キーワードで構成されている。キーワードが1個しか現れていない記事は「ビデオ」や「インフォメーション」という1つのキーワードのみであり内容のある文書とは言えないが、本論文への影響は少ないと考え、含めた。

5. 野球チームの条件付き確率分布

5.1. 野球チーム名について

本研究では、野球チーム名は日本のプロ野球球団名を指す。対象とした新聞記事が91年から97年であるため、当時の球団名を用いた。具体的には、セ・リーグには「巨人」「中日」「阪神」「ヤクルト」「横浜」「広島」、パ・リーグには「西武」「ダイエー」「日本ハム」「オリックス」「ロッテ」「近鉄」という文字列を用いた。

5.2. 各チームの cf と df

それぞれ野球チーム名のキーワードに対する df と cf を、セ・リーグは表 1に、パ・リーグは表 2に示す。セ・リーグについては cf にバラつきが見られる。特に別に地名としての意味を持つ「阪神」「横浜」「広島」は cf が高く、他の球団の 3,4 倍になっている。一方パ・リーグは全て企業名がつけられており、cf は 1 万回前後となっている。

また、野球チーム名は他にも「巨人」であれば「ジャイアンツ」、「西武」であれば「ライオンズ」という別名も存在するが、こちらは cf が低いために十分なデータを取ることができなかった。これら別名についての cf と df を表 3 に示す。

表 1 セ・リーグ球団名の cf と df

w	cf(w)	df(w)	cf/df
巨人	13348	4508	2.96
中日	8912	3331	2.68
阪神	42474	23962	1.77
ヤクルト	10794	3566	3.03
横浜	30443	16851	1.81
広島	37664	17097	2.20

表 2 パ・リーグ球団名の cf と df

w	cf(w)	df(w)	cf/df
西武	13667	4453	3.07
ダイエー	11885	4342	2.74
日本ハム	6713	2283	2.94
オリックス	9414	3371	2.79
ロッテ	7260	1571	4.62
近鉄	9799	2080	4.71

表 3 プロ野球球団別名の cf と df

w	cf(w)	df(w)	cf/df
ジャイアンツ	1159	951	1.22
ドラゴンズ	108	93	1.16
タイガース	1141	886	1.29
ライオンズ	243	187	1.30
ホークス	190	134	1.42
ブルーウェーブ	173	148	1.17

5.3. 各チームの条件付き確率グラフ

それぞれ野球チーム名のキーワードに対する繰り返しの条件付き確率を、セ・リーグは図 2に、パ・リーグは図 3に示す。セ・リーグ、パ・リーグ共に見られる特徴として、Pw(4|3)や Pw(5|4)から右肩下がりのグラフになっていることが挙げられる。特にパ・リーグは顕著であり、西武以外の球団名は、全て Pw(2|1)から P(4|3)にかけて値が上昇し、その後値が下がり P(7|6)で横ばいとなっている。セ・リーグはパ・リーグほどはっきりとは見られないが、値が一度上がってから下がるという特徴はパ・リーグと変わらない。

更に、これら野球チーム名のグラフでは k が大きくなるにつれ P(k+1|k)が減少しているため、図 1で示したような x 軸と平行な直線になる Katz K mixture モデルには合致しないことがわかる。

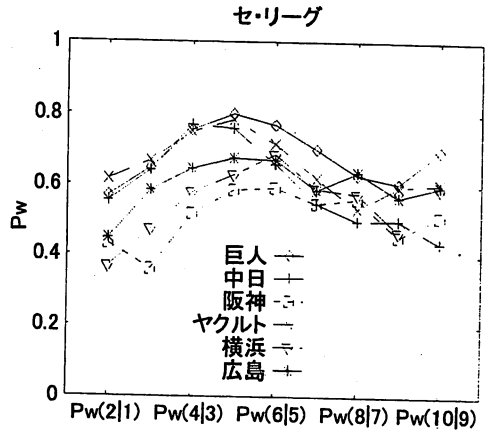

図 2 セ・リーグ球団の条件付き確率

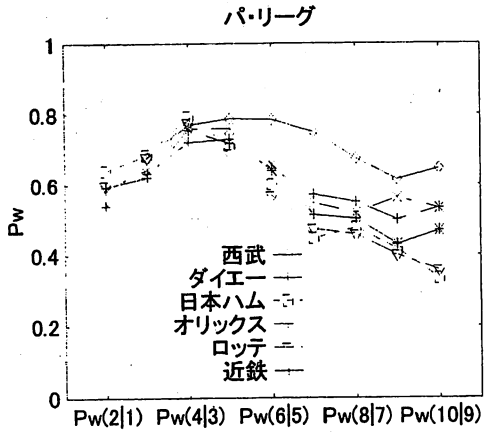

図 3 パ・リーグ球団の条件付き確率

6. 他のキーワードとの比較

本研究で用いている野球チーム名は、チーム名以外の意味を持つ。例えば「横浜」「広島」「阪神」というのは地名という意味も含まれており、野球チームとは関係なく使われることがある。また、その他のチーム名のほとんどは企業名として使われている。

そこで、これら野球チーム名として以外の意味によって図 2、図 3のような結果が現れているのか、ということについて検証をするため、他の地名や企業名、また他のスポーツで現れるキーワードについての検証を行った。

6.1. 他の地名との比較

初めに、他の地名との比較を行う。プロ野球チームでは「横浜」「広島」「阪神」が地名に当たる。比較対象としたのは、「札幌」「仙台」「東京」「大阪」「神戸」「名古屋」という主要都市である。

これら地名に対する cf と df を表 4に示す。「東京」「大阪」は頻出単語であるため cf が大きくなっているが、他の 4 都市に関しては「東京」「大阪」と比べて「阪神」「横浜」「広島」と同程度の cf を示している。

これら地名に対する繰り返しの条件付き確率を図 4に示す。このグラフからは、どの地名も野球チーム名として使われている「阪神」「横浜」「広島」とは違った性質を持っていることがわかる。初め Pw(2|1)から Pw(4|3)まで値が上昇しているのはどのキーワードも変わらないが、その後野球チームのキーワードに見られた値の低下を見ることはできず、平坦、もしくは少し上昇するといった傾向が見られる。

この結果から、地名はプロ野球チームと違う性質を持っている事がわかり、地名という意味については影響がないと考えられる。

表 4 主な地名の cf と df

w	cf(w)	df(w)	cf/df
札幌	10232	6536	1.57
仙台	8090	5121	1.58
東京	248594	144747	1.72
大阪	284416	147789	1.92
神戸	50424	24457	2.06
名古屋	19182	11617	1.65

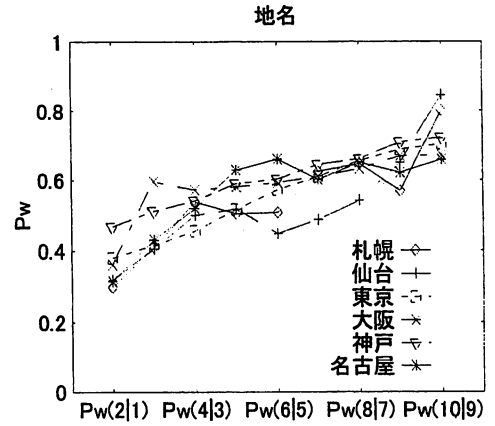

図 4 地名の条件付き確率

6.2. 他の企業名との比較

次に、他の企業名との比較を行った。野球チーム名のうちで企業名が使われているのは、「中日」「ヤクルト」とパ・リーグの全球団が当てはまる。ここで挙げる企業名は、「ロッテ」に対して「グリコ」「森永」「明治」、「中日」に対して「読売」「朝日」、「西武」に対して「小田急」である。

これら企業名に対する cf と df を、表 5に示す。本研究の対象が新聞記事であるために広告としての役割は果たさないからか、企業名の cf はあまり高くない。また、ここで挙げた企業の多くはプロスポーツに参加していないため、企業名を持つ野球チーム名と比べ cf が小さい。

これら企業名に対する繰り返しの条件付き確率を図 5に示す。cf が低いため、Pw(k+1|k)の k が高い領域において推定値が不安定なキーワードが現れてしまった。また、このようなキ

ーワードを除くと値の下がっているキーワードは他に無く、野球チーム名と似た性質のキーワードを見つけることはできなかった。

この結果から、プロスポーツに関係しない企業名はプロ野球チームとは違った性質を持っており、企業名という意味については影響が無いと考えられる。

表 5　主な企業名の cf と df

w	cf(w)	df(w)	cf/df
グリコ	631	268	2.35
森永	673	412	1.63
読売	1725	1117	1.54
朝日	3330	2226	1.50
明治	8266	5871	1.41
小田急	1124	777	1.45

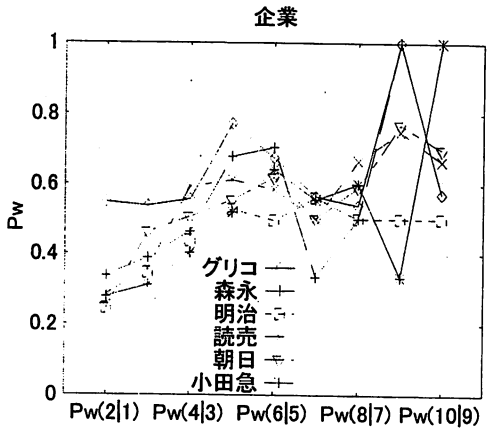

図 5　企業名の条件付き確率

6.3. 他のスポーツとの比較

最後に、他のスポーツとの比較を行う。ここでは代表的なプロスポーツとしてサッカーとラグビーについて調べる。サッカーについては「清水」「川崎」「市原」「鹿島」と、別名として「エスパルス」「ヴェルディ」というキーワード、ラグビーについては「サントリー」「トヨタ」「神戸製鋼」「ＮＥＣ」というキーワードを選択した。サッカーのクラブ名として挙げた「清水」「川崎」「市原」「鹿島」はそれぞれ地名を表しており、野球チーム名であれば「阪神」「横浜」「広島」といったチーム名と同じ性質があるとも考えられる。また、ラグビーのチーム名として挙げたキーワードはそれぞれ企業名として使われているため、野球チーム名の中でも企業名として使われているキーワードと同じような性質があると考えられる。

これらサッカークラブ名に対する cf、df を表 6に、ラグビーのチーム名に対する cf、df を表 7に示す。サッカーのチーム名のうち、「エスパルス」「ヴェルディ」は野球チームの別名と同様低い cf であった。この結果から、新聞記事においてサッカーチームは地名で書かれていることがわかる。ラグビーチームについては、野球チームで使われている企業名と比べ cf が小さい。また表 5と比較すると、企業名に関してはプロスポーツに参加することで、新聞記事に現れることが多くなるということがわかった。

サッカークラブ名とラグビーチームに対する繰り返しの条件付き確率を、それぞれ図 6と図 7に示す。

野球チーム名として使われているキーワードはどれも同じような軌道を持っていたことに対し、サッカークラブ名でも同じ傾向が見られた。ここで挙げたクラブ名に関しては、一様に同じ軌道を描いているといえる。野球チーム名と比べ Pw(2|1)から Pw(4|3)まで上昇している点については同じだが、Pw(5|4)以降は値が下がらず x 軸と平行という軌道になっているため、やはり野球チーム名とは異なる特徴を持っていることがわかった。また、サッカークラブ名の別名である「エスパルス」「ヴェルディ」については、cf が小さいために有意なグラフを得ることはできなかった。ラグビーチーム名のグラフには、サッカークラブ名と同じ特徴が現れた。ラグビーチーム名についても Pw(2|1)から値が上昇し、Pw(5|4)からは x 軸と平行な軌道となっている。

これらの結果から、野球チームの条件付き確率はスポーツとも関係ないことがわかった。また、サッカークラブ名とラグビーチーム名には同じような傾向が見られたため、スポーツに関係しているキーワードに関しては同じ軌道を描くという可能性も示された。

表 6　主なサッカークラブ名の cf と df

w	cf(w)	df(w)	cf/df
清水	13658	7605	1.80
川崎	15962	8254	1.93
市原	4717	2048	2.30
鹿島	8327	3372	2.47
エスパルス	490	393	1.25
ヴェルディ	466	284	1.64

表 7 主なラグビーチーム名の cf と df

w	cf(w)	df(w)	cf/df
サントリー	4504	2411	1.87
トヨタ	4218	2146	1.97
神戸製鋼	1668	876	1.90
NEC	5681	2865	1.98

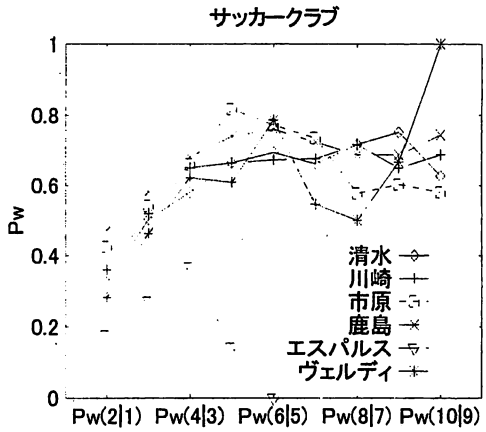

図 6 サッカークラブ名の条件付き確率

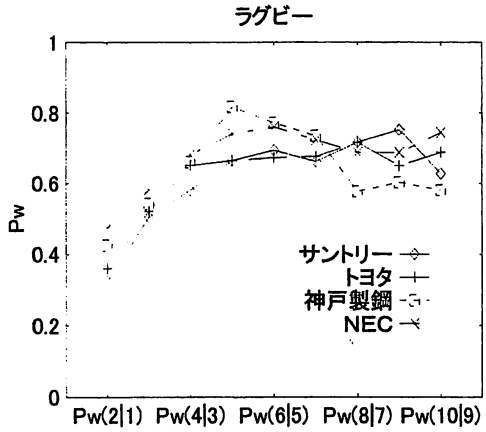

図 7 ラグビーチーム名の条件付き確率

7. まとめ

本論文では、野球チーム名が Katz K mixture モデルと合致せず、繰り返しの条件付き確率に関する特異な分布を持つことついて述べた。

他の地名、企業名、スポーツチーム名についても同様に繰り返しの条件付き確率を調べ、プロ野球チーム名のみに特徴が現れていることを示した。

今回調査した地名、企業名、スポーツチーム名については k が増えた時に Pw(k+1|k)の値が上昇するという特徴を持っているが、野球チーム名だけは全て値が下がっている。この点から、Katz K mixture モデルに合致しないばかりか、合致しないキーワードの中でも更に特異であると考えられる。

繰り返しが多い回数でのモデルの提案を行うために、野球チームと同じような特徴を持つキーワードを探し、なぜこのような特徴が現れるか、ということを検証したい。

文 献

[1] Katz, S. M. "Distribution of content words and phrases in text and language modeling." Natural Language Engineering, vol.2(1), 15-59. 1996.

[2] Yinghui Xu, Kyoji Umemura. "Improvements of Katz K Mixture Model" 自然言語処理 , Vol.12 , No.5 , Oct.2005.

[3] 毎日新聞社. 毎日新聞データ 91,92,93,94,95,96,97 年版.

[4] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム『茶筌』version 2.3.3 使用説明書, Aug,2003.

[5] 梅村恭司, 真田亜希子. "文字列を k 回以上含む文書数の計数アルゴリズム" 自然言語処理 , Vol.9 , No.5 , Oct.2002.