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

2007－NL－180（25）
2007／7／25

テンプレートを構成する名詞のKatzモデルによる抽出の試み

藤原 大輔† 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: †fuji@ss.ics.tut.ac.jp, †††umemura@tutics.tut.ac.jp

あらまし 文書の雛形をテンプレートと呼ぶが，あるテンプレート内で使用される単語をテンプレートの形を知ることなく抽出するという問題を扱う．単語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，単語が文書中で繰り返し出現する条件付確率は減衰係数によって決められると仮定している．本研究では，このKatz K mixtureモデルに従わない固有名詞が持つ特徴とテンプレートの関係について分析し，その結果，モデルに合致しないものがテンプレート内で使用される単語の候補となり得ることが分かった．

キーワード Katzモデル 統計的言語処理 テンプレート 単語頻度 固有名詞

Extracting Nouns that Constitute Templates by the Katz Model

Daisuke FUJIHARA†, Akihiro TAKASE†*, and Kyoji UMEMURA†

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

Abstract A template is a fixed format of certain documents. We deal here with the problem of extraction words used in templates without knowing form of the templates. The Katz K mixture model is well known as a distribution model of keywords. In this model, basic assumption is that the conditional probabilities of repeats for a given word are determined by a decay factor. In this study, we analyze relations of a template and proper nouns which do not obey the Katz K mixture model. As a result, we have found that the Katz model is useful to detect nouns that consitute templates.

Key words the Katz K mixture model, statistical natural language processing, template, term frequency, proper noun

1. はじめに

近年のインターネットの普及により，我々が利用できる情報の量は急激に増加している．特に各種ニュースサイトやブログは更新頻度が高く，この情報量の増加を担う一因である．

統計的言語処理は情報量の増加に強いことが知られている．しかし，文書の雛形であるテンプレートの存在が統計的言語処理の性能を低下させる場合がある．例えば，テンプレートとして用いられる文字列は一般に有益な情報ではないため，検索対象から除外したほうが良い．既にテンプレートとして共通に現れる文字列を除去したインデックスを用いることで検索性能を上げるという研究が存在する [1].

事前にテンプレートのフォーマット情報を持たずにテンプレートを発見する方法として部分文字列増幅法 [2] が提案されている．この方法は，コーパス内の情報に対して部分文字列を作成し，その部分文字列に対する頻度の情報を用いてテンプレートを抽出するというものである．しかしこの方法はテンプレートのフォーマットを見つける研究であり，その内容である単語には触れられていない．

一方，本研究が対象とするのは，テンプレートのフォーマットを全く考えることなく，テンプレート内に含まれる情報そのものを抽出するという問題である．もしテンプレートに関する情報を利用することなく，単語に関する統計情報のみを用いて特定することができれば，フォーマットが統一されていないデータにおいて利用できるだろう．

本研究では，テンプレートを構成する単語を抽出するための手がかりとして，Katz K mixtureモデル [3] に従わないものを候補とすることを提案する．そしてこれらの単語とテンプレー

(注*)：現 NTT 東日本

トとの関係について分析する．

2. 記号の定義

本論文で用いる記号の定義を以下に示す．なお，N 以外の記号は，単語 w にが前提として定まっているものとし，引数に含めないものとする．

N : コーパス数

cf : 単語 w のコーパス中の出現回数

$df(k)$: 単語 w が k 回以上出現した文書数

$k=1$ のときは単に df と表す

$tf(d)$: 文書 d における単語 w の出現回数

$cdf(k)$: 単語 w が k 回以上出現した累積の文書数

$$cdf(k) = \sum_{i=1}^{k} df(i)$$

$Pw(k+1|k)$: 単語 w における繰り返しの条件付確率

$$Pw(k+1|k) \equiv P\left(tf(D) \geqq k+1 | tf(D) \geqq k\right)$$

ここで，D 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 モデルは単語の繰り返し条件付確率について，「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．ここでは，2 パラメータモデルについてその定義を示す．

3.2 定　　義

Katz K mixture モデルは，「文書中に単語 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．

$$\begin{aligned} P_{Katz} &= P\left(tf(X)=k\right) \\ &= (1-\alpha)\delta_{k,0} + \frac{\alpha}{\beta+1}\left(\frac{\beta}{\beta+1}\right)^{k} \\ &\text{where } \delta_{k,0} = \begin{cases} 1 & \text{iff} \quad k=0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

このうち，α と β は平均 λ や逆文書頻度 IDF から以下のように算出される．

$$\hat{\lambda} = \frac{cf}{N} \tag{2}$$

$$I\hat{D}F = \log_2 \frac{N}{df} \tag{3}$$

$$\hat{\alpha} = \frac{\hat{\lambda}}{\hat{\beta}} = \frac{cf}{N}\frac{df}{cf-df} = \frac{df}{cdf(2)} \tag{4}$$

$$\hat{\beta} = \hat{\lambda}2^{I\hat{D}F} - 1 = \frac{cf-df}{df} = \frac{cdf(2)}{df} \tag{5}$$

これより，$P_{Katz}(k+1)/P_{Katz}(k)$ は 2 パラメータモデルの減衰係数 $\hat{\beta}/(\hat{\beta}+1) = cdf(2)/cf$ によって決定される．

3.3 Kats K mixture モデル例

Katz K mixture 2 パラメータモデルが示す単語の繰り返し条件付確率のグラフを図 1，図 2，図 3 に示す．

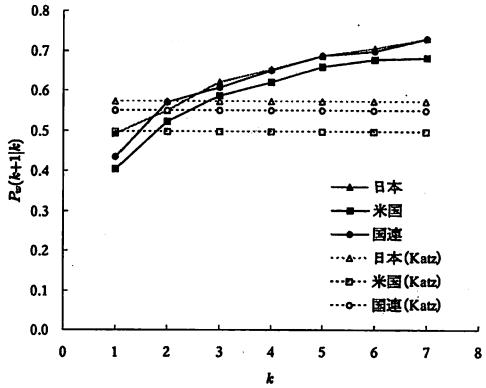

図 1　固有名詞の場合における比較例

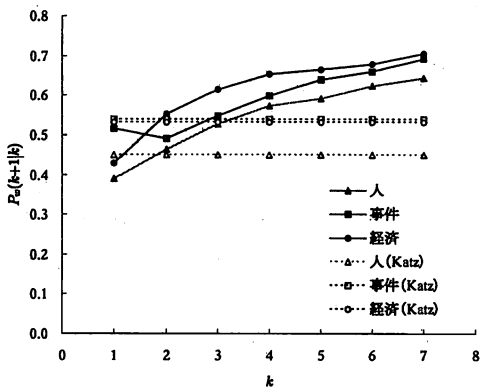

図 2　一般名詞における比較例

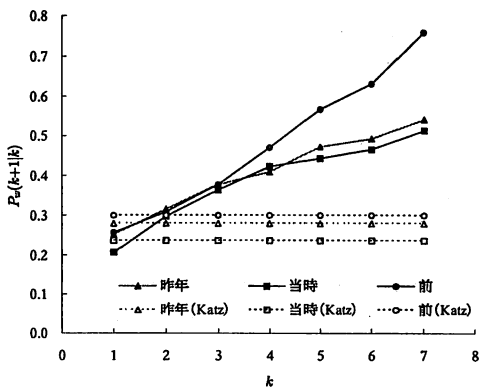

図 3　時制を表す名詞における比較例

これらのグラフは，各単語に対してモデルの推定値と実際の値の推移を表したものである．なお，繰り返しの条件付確率は $\hat{P}_w(k+1|k) = df(k+1)/df(k)$ で推定される．

2パラメータモデルは，x 軸に水平となる．実際の値は多くの場合，水平か小さな傾きで増加する傾向が見られる．単語によっては図3に示した「前」のような，大きな傾きで増加するものもある．Katz モデルは英語に対して提案されたモデルであり，Xu ら [4] は Katz K mixture モデルのパラメータを動的に決めるように改良し，高瀬 [5] はこれらのモデルが日本語に対しても有効であることを示した．我々はこの分析から，単語の繰り返し条件付確率が減少するケースに興味を持った．

4. 実験の設定

以下の実験では，Katz K mixture モデルに従わない単語を抽出し，それらの中にテンプレートを構成する要素が含まれていることを示す．本実験で使用したコーパスは，毎日新聞の91年から97年までの2,550日分（毎年1月2日は休刊）の全731,548記事を集めたもの [6] である．記事の内容は，政治，経済，スポーツ，社会，対談などさまざま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処理を順に実行した．

1. MeCab0.95，辞書 ipadic2.7.0 を用いた形態素解析
2. 名詞および未知語を抽出
3. 各名詞に対し，$df(k)$ を計測
4. 名詞の中から，$df(5)$ が10以上のものを抽出
5. 辞書の定義と目視により，固有名詞を抽出
6. 各単語に対し，$P_w(k+1|k)$ を算出

ここで，固有名詞と判定したものは，MeCab の出力が固有名詞となったもの，MeCab が未知語と出力したものから目視で判断したものに加え，英語表記では固有名詞となる「1月」「2月」や「月曜」「火曜」などを加えたものである．また，一般名詞となる場合と固有名詞となる場合がある単語は文脈に関わらず固有名詞と考えて出現頻度をカウントした．この結果，最終的抽出した固有名詞は3,649語であった．この語が以下の実験の対象となる．

5. 実験：Katz モデルに従わない単語の抽出

図1から図3に示したように，一般的な単語の場合には k が増加するにつれて $P_w(k+1|k)$ も増加する傾向がある．このことから，$P_w(k+1|k)$ が減少するものをすべて取り出し，どのようなものが含まれるか分析した．

ここでは，Katz モデルに従わない条件確率 $P_w(k+1|k)$ の推移をもつ単語を，ある k の範囲で単調増加しないものと定義した．k の範囲は $[1,7]$ とした．本来ならば，統計的手法を用いて検定する必要があるが，今回は簡単化のためにこのように定義した．

さらに特徴を分類するため，$\underset{k}{\mathrm{argmax}} P_w(k+1|k)$ について着目する．なお，実験対象の名詞3,649語に対し，k の範囲 $[1,6]$ で最大値を持つものは2,264語であった．

Katz モデルに従わない単語の一部とその単語の繰り返し条件確率の推移を分野別にして以下に示す．

表1 プロ野球球団名と呼称

リーグ	球団名	呼称
セントラルリーグ	中日ドラゴンズ	中日
	阪神タイガース	阪神
	広島東洋カープ	広島
	ヤクルトスワローズ	ヤクルト
	横浜大洋ホエールズ	大洋
	横浜ベイスターズ	横浜
	読売ジャイアンツ	巨人
パシフィックリーグ	大阪近鉄バッファローズ	近鉄
	オリックスブルーウエーブ	オリックス
	西武ライオンズ	西武
	千葉ロッテマリーンズ	ロッテ
	日本ハムファイターズ	日本ハム
	福岡ダイエーホークス	ダイエー
	ロッテオリオンズ	ロッテ

表2 野球球団呼称の条件付確率最大の k

呼称	$\underset{k}{\mathrm{argmax}} P_w(k+1|k)$	呼称	$\underset{k}{\mathrm{argmax}} P_w(k+1|k)$
中日	3	近鉄	4
阪神	5	オリックス	4
広島	4	西武	4
ヤクルト	4	ロッテ	3
大洋	4	日本ハム	3
横浜	5	ダイエー	3
巨人	4		

5.1 プロ野球球団名の呼称

91年から97年の間に存在した日本のプロ野球球団名と，最も多く使用される呼称を表1に示す（ロッテオリオンズは91年まで，横浜大洋ホエールズは92年まで，千葉ロッテマリーンズは92年から，横浜ベイスターズは93年からのプロ野球球団）．今回は呼称13種を対象とした．セントラルリーグ球団の条件付確率の推移を図4に，パシフィックリーグ球団の場合を図5にそれぞれ示す．また，各呼称の $\underset{k}{\mathrm{argmax}} P_w(k+1|k)$ をまとめたものを表2に示す．

表2から，$P_w(k+1|k)$ が $3 \leqq k \leqq 5$ の範囲で最大となっていることが分かる．$P_w(k+1|k)$ が大きいということは，単語 w が k 回出現したときに $k+1$ 回出現する確率が高いということを表す．

また，図4，図5を見ると，全体的な傾向も近いことが分かる．特にパシフィックリーグの西武を除く5球団は，$P_w(k+1|k)$ の値，増加・減少の傾きが類似している．

ここで注目すべき事象として，球団呼称は毎日新聞コーパス中において実際にテンプレート中に使用されていたということである．このテンプレートは，プロ野球の試合結果を示すためのものであった．

5.2 サッカーチームの地名

Jリーグのチーム名について実験する．プロ野球球団と同様，最も使用される呼称である地名に対して実験した．ただし野球チームの場合と異なり，91年から97年までに存在したチーム

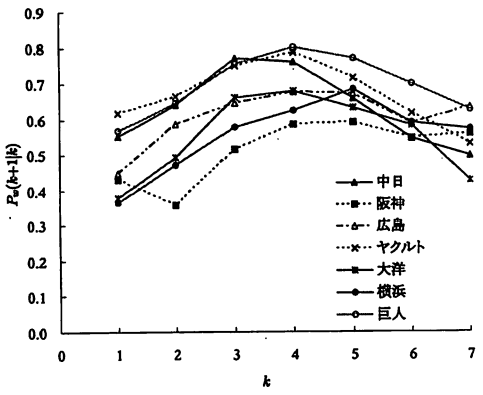

図 4　セ・リーグの球団呼称の条件付確率

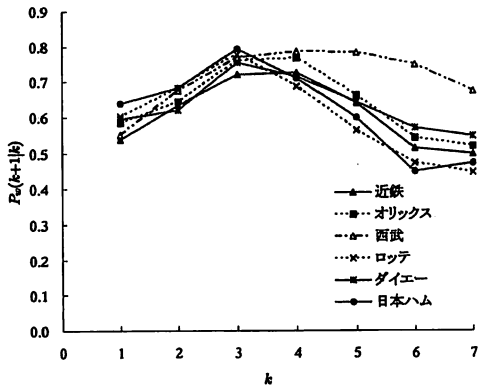

図 5　パ・リーグの球団呼称の条件付確率

表 3　対象とする J リーグチーム名と条件付確率最大の k

チーム名	$\underset{k}{\mathrm{argmax}} P_w(k+1\|k)$	チーム名	$\underset{k}{\mathrm{argmax}} P_w(k+1\|k)$
市原	4	鹿島	5
磐田	5	横浜フ	3
浦和	5	横浜マ	4

の中から Katz モデルに従わないものを選んだ．選出した単語と $\underset{k}{\mathrm{argmax}} P_w(k+1|k)$ をまとめたものを表 3 に示す．また，条件付確率の推移を図 6 に示す．

表 3 から，$P_w(k+1|k)$ が $3 \leqq k \leqq 5$ の範囲で最大となっていることが分かる．これはプロ野球球団の場合と同様である．また，図 6 を見ると，全体的な傾向も近いことが分かる．

また J リーグチームの場合もプロ野球球団同様，試合結果を示すための記事に存在し，その記事はテンプレートが使用されていた．

5.3　曜　　日

曜日を表す単語を対象に実験した．対象の単語と $\underset{k}{\mathrm{argmax}} P_w(k+1|k)$ をまとめたものを表 4 に示す．ここで，「月曜」「火曜」「水曜」「木曜」は $\underset{k}{\mathrm{argmax}} P_w(k+1|k)$ が $3 \leqq k \leqq 5$

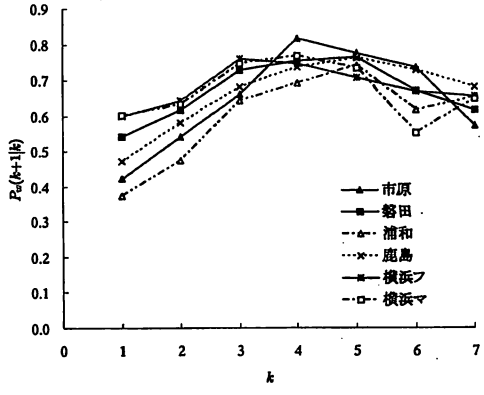

図 6　J リーグチーム名の条件付確率

表 4　曜日の条件付確率最大の k

曜日	$\underset{k}{\mathrm{argmax}} P_w(k+1\|k)$	曜日	$\underset{k}{\mathrm{argmax}} P_w(k+1\|k)$
月曜	3	金曜	7
火曜	5	土曜	7
水曜	4	日曜	7
木曜	3		

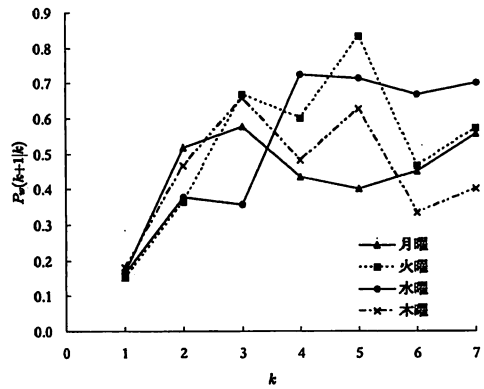

図 7　月曜～木曜の条件付確率

となっているが，「金曜」「土曜」「日曜」は $\underset{k}{\mathrm{argmax}} P_w(k+1|k)$ が 7 となっている．それぞれの条件付確率の推移を図 7，図 8 に示す．

図 7 を見ると，「火曜」「木曜」は $k = 3, 5$ において $P_w(k+1|k)$ が大きい傾向あり．「木曜」は $k = 3$ で最大となっている．しかし，「水曜」は $k = 4$ で最大ではあるが，その後の増減は小さい．

また図 8 を見ると，「金曜」は $P_w(5|4)$ が $P_w(4|3)$，$P_w(6|5)$ に比べて高くなっている一方，「土曜」は $k = 5$ で減少しているがその幅は小さく，「日曜」に関しては単調増加である．

したがって，曜日の一部には Katz モデルに従っていないものが見られるが，すべての曜日では無いためテンプレートが影響しているとは言えず，なぜ Katz モデルに従っていないかの

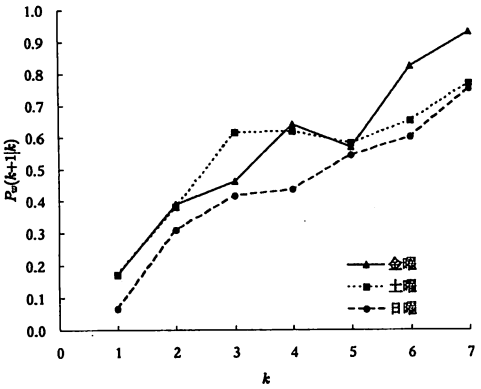

図 8　金曜～日曜の条件付確率

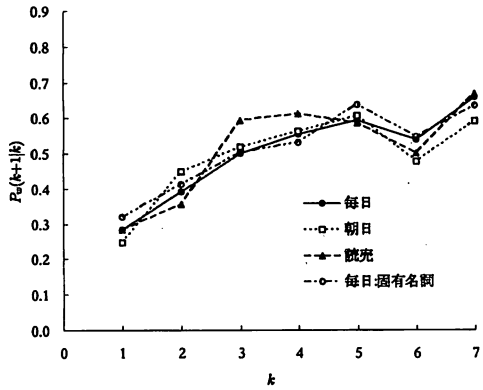

図 9　新聞名の条件付確率

理由は不明である.

5.4 新 聞 名

全国紙の新聞名「毎日」「朝日」「読売」を対象に実験した. 条件付確率の推移を図 9 に示す. さらに比較のため,「毎日」の中で MeCab が前後の文脈から固有名詞と判断したものの繰り返し条件付確率も図示した.

なお MeCab による形態素解析では, 上記の単語は「毎日新聞」「毎日放送」「テレビ朝日」「朝日新聞」「読売新聞」などとは別のものと判定される.

$\underset{k}{\mathrm{argmax}} P_w(k+1|k)$ は,「朝日」が 4,「毎日」「読売」が 7 となっているが, $P_w(7|6)$ が $P_w(6|5)$, $P_w(8|7)$ に比べて低くなっていることが共通点として挙げられる. また「毎日」は一般名詞としても使用されるが, 固有名詞に限定したグラフを見ると, 同様に $P_w(7|6)$ が低くなっていることが分かる. したがって, この特徴は新聞名によるものだと推定できる.

6. テンプレートと候補の分析

5.1 節, 5.2 節の実験にも示したように, 野球の結果の記事, サッカーの結果の記事において, 実際にテンプレートがあることが分かっている.

そして実験から, プロ野球の球団呼称はすべて Katz K mixture モデルの分布から外れているし, J リーグのチームの一部も Katz K mixture モデルの分布から外れている.

したがって, Katz K mixture モデルに従わない単語は, テンプレート内で使用される単語の候補となりうる.

しかしながら, Katz K mixture モデルに従わない単語が, すべてテンプレート内で使用されるとは言い切れない.

なぜなら, 繰り返し条件付確率が単調増加しない単語すべてについて, 明らかなテンプレートが存在するとまで言えないからである. この原因として, 単調増加しないという基準では偶然のケースも含み得ることがある. また, 新聞記事内には固有名詞が繰り返し現れるテンプレートばかりではないということも考えられる. 実験の結果から, 今回の方法がテンプレートのある側面を反映していると考えられるので, テンプレートであるという判定基準の検討と, 他種のコーパスに対する実験が必要となる.

7. ま と め

テンプレートのフォーマットを全く考えることなく, テンプレート内に含まれる単語を抽出するという問題に対して, Katz K mixture モデルに従わない単語がその候補となることを示した.

今後の課題としては, 以下のものが挙げられる.

- ひとつのドキュメント中に, 単語が複数回出現するようなテンプレートにしか効果が無い

これは, 本手法が単語の繰り返しの情報を利用していることに起因していて, ある単語がひとつのドキュメントに 1 回しか出現しないテンプレートであれば抽出は難しい.

- テンプレートかそうでないかを決定する基準

これは, テンプレートによってピークとなる k の値が異なることと, テンプレートに使用されない単語でも, Katz モデルに従わないものが存在することが要因である.

謝辞　この研究は住友電工情報システム (株) との共同研究の成果であり, 戦略的情報通信開発推進制度 (SCOPE) の課題「実空間情報処理のためのインターユビキタスネットワークの研究」に使用する予定です.

文　献

[1] Z. Bar-Yossef and S. Rajagopalan: "Template detection via data mining and its applications", Proc of the 11th international conference on World Wide Web (2002).

[2] 池田, 山田, 廣川: "部分文字列増幅法による共通パターン発見アルゴリズム", 情報処理学会論文誌, **46**, 2, pp. 56–66 (2005).

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

[4] Y. Xu and K. Umemura: "Improvements of katz k mizture model", 自然言語処理, **vol.12**, No.5 (2005).

[5] 高瀬: "キーワードの異常出現検定のための統計モデルに関する研究", 修士論文, 豊橋技術科学大学 (2006).

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