

キーワード抽出を実現する文書頻度分析

武田善行 梅村恭司

豊橋技術科学大学 情報工学系

441-8580 豊橋市天伯町雲雀ヶ丘 1-1

TEL(+81)532-47-0111(ex.5430)

take@ss.ics.tut.ac.jp umemura@tutics.tut.ac.jp

反復度とは文書においてある部分文字列が 1 回以上出現するという条件でその部分文字列が 2 回以上出現する度合いである．本論文では英語において観測されているキーワードの反復出現が日本語においても観測できることを確かめた．英語同様に，キーワードの反復度はその頻度に対して無相関であった．一方，ランダムに切り出された文字列の反復度はばらついていた．この分析を日本語論文抄録と数年の日本語新聞記事で行い，反復度がキーワード境界の特定が可能な情報を持つことを示した．

語分割，反復度，反復出現，多言語，キーワード抽出

Document Frequency Analysis which Realizes Keyword Extraction

Yoshiyuki Takeda Kyoji Umemura

Dept. of Information and Computer Sciences, Toyohashi University of Technology

Tempaku, Toyohashi, Aichi, 441-8580, Japan,

TEL: (+81)532-47-0111(ex.5430)

take@ss.ics.tut.ac.jp umemura@tutics.tut.ac.jp

Adaptation is the degree in which a substring appears twice or more, when it appears once or more in a document. Adaptation of the keyword has been observed in English. Similarly, it is observed in Japanese and Chinese. We have observed that adaptation of a keyword tends to have no correlation with just like English. On the other hand, the estimated value varies in strings that are selected at random. We analyzed adaptation using newspaper article of several years and technical abstracts. We have tried to extract keywords using the difference of this distribution. We show that adaptation contains the information with which keyword boundaries are obtained.

Word segmentation, Adaptation, Repetitive occurrence, Multilingual, Keyword extraction

1. はじめに

我々は単語の反復出現の度合を捉えた特徴量 [1] について、文献[1]より一般的な条件で分析を行い、その結果、この特徴量が日本語のような単語の境界についての情報を持たない言語において、キーワードを抽出するための特徴量として有効であることを報告した [2]。文献[2]より日本語に関わる部分を整理したのが本論文である。

常に新しい概念やそれを示す語が出現する状況において、キーワードのリストを構築し、維持するコストは高い。電子化された大量のコーパスの入手が以前より容易となった現状では、コーパスベースの統計的手法が有効であるとされているが、日本語の文章は区切り情報を持たないため、コーパス以外の文法知識や語の情報が必要となる。

統計的手法を用いる場合、キーワードを特徴付ける統計量として語の出現頻度が多く用いられているが、この頻度に限らず、文書集合や分野毎の分布の偏り[3]を使うことが考えられる。文献[1]では、単語の反復出現の度合を用いた特徴量を提案している。本論文では文献[1]で提案された手法について日本語のコーパスでの振る舞いを分析した。結果として、英単語における報告と同様に、反復度が語彙に強く依存するという結果を得た。

また、文献[1]における英単語での分類調査に対して、より一般的な状況を考え、日本語のコーパスにおいて単語の区切り情報を使わずに分析を行った。具体的には、長さ n において、 2^{n-1} 通りのすべての分割を考え、それによってできる $n(n-1)$ 通りの部分文字列についての頻度分析を行った。この分析の結果、反復度がキーワードの境界を特定するために有効であるという結果を得た。この結果を基に、語分割の必要な日本語においてキーワードを抽出するための尺度として反復度が有効であることを報告する。

2. 反復度の定義

多くの語は文書に繰り返し出現し、一度出現した語は再び出現する傾向にある。そして、その度合いは語の意味に関わる値であることが報告されている[1]。文献[1]では、語の持つ特徴量として他にも提案しているが、本論文では次に示す特徴量に注目した。

定義 2.1 反復度

ここでの全事象は文書の出現とする。語を w 、文書が語 w を (1 回以上) 含む事象を $e_1(w)$ 、文書が語 w を 2 回以上含む事象を $e_2(w)$ とすると、反復度は次のよう

に定義される。

$$\begin{aligned} \text{adaptation}(w) &= p(e_2(w) | e_1(w)) = \frac{p(e_2(w) \wedge e_1(w))}{p(e_1(w))} \\ &= p(e_2(w)) / p(e_1(w)) \end{aligned}$$

反復度は、ある文書に単語 w が 1 回以上含まれていることを条件とした時にある文書に単語 w が 2 回以上含まれる条件付き確率である。語が文書に出現する確率と語が文書に 2 回以上出現する確率は、コーパス全体で考えると文書頻度を用いて推定することができる。コーパス全体である語 w を含む文書の数を df 、2 回以上含む文書の数を df_2 とすると、反復度は以下のように推定することができる。

$$\text{adaptation}(w) = p(e_2(w)) / p(e_1(w)) \approx df_2 / df$$

また、反復度の持つ特徴として次のようなことが報告されている[1]。

- 自立語は df_2 / df は高く、付属語の df_2 / df は低い。
- df / N と df_2 / df とは経験的に一桁か二桁違う。ただし、 N はコーパス全体での文書数とする。
- df_2 / df は df / N に対して無相関である。

本論文で注目する反復度の特徴は、ある文書において一度出現した語がもう一度繰り返される度合いが語の種類と密接な関係を持っている点である。統計的独立を考えた場合の反復度に対して、自立語の反復度は非常に高い。これは、統計量でしかない df_2 / df に対する言語的な意味付けである。ここで量られるのは文書において語が繰り返し出現する度合いであるため、本論文では df_2 / df を反復度と呼ぶ。

3. 反復度と頻度の関係

本論文では文献[1]での分析に対して、より一般的な状況を考え、単語における反復度の振る舞いに限らず、任意に切り出した部分文字列における反復度の振る舞いに興味を持った。言語により、語もしくは重要な意味を持つ単位の境界は明確でない。たとえば、日本語や中国語の文章は語を区切る空白がない。語の境界が明確でないならば、語でないものを含めて分析する価値があると本論文では考えた。文献[1]での分析は語彙による分類調査であるため、任意に切り出した部分文字列における反復度の振る舞いは不明である。本論文では、任意に切り出した部分文字列について、反復度の分布を分析する。

分析対象となるコーパスとして NTCIR テストコレクション[4]、CD-毎日新聞 91～97 版[5]を用いた。それぞれの実体は、論文抄録と新聞記事である。それぞれ

表 1 分析対象となるコーパス

コーパス	件数	内容
NTCIR1	332,921	NTCIR-1 J コレクション (学会発表データベースの一部)
NTCIR2G	116,177	NTCIR-2 J コレクション (学会発表データベースの一部)
NTCIR2K	287,071	NTCIR-2 J コレクション (科学研究費補助金研究成果概要データベースの一部)
MAI91	91,200	CD-毎日新聞 91 版
MAI92	101,468	CD-毎日新聞 92 版
MAI93	91,774	CD-毎日新聞 93 版
MAI94	101,058	CD-毎日新聞 94 版
MAI95	111,497	CD-毎日新聞 95 版
MAI96	114,729	CD-毎日新聞 96 版
MAI97	119,836	CD-毎日新聞 97 版

のコーパスについて詳細を表 1 に示す。

NTCIR1, NTCIR2G, NTCIR2K は情報検索問題における性能評価を主眼としたテストコレクションであり、文書と質問の 2 種類の文書からなる。それぞれの文書や質問にはその文章の著者が付けたキーワードが付属している。本論文ではこれをランダムに抽出したものをそれぞれのコーパスにおけるキーワードの典型的なものとして分析を行った。また、MAI91 ~ MAI97 にはこのようなキーワードが付属していなかったため、NTCIR1 に付属していたキーワードを用いた。確認のため NTCIR2G, NTCIR2K のキーワードを用いて実験を行ったが、結果に差異はなかった。

それぞれのコーパスについて、図 1 の上側に任意に切り出した部分文字列を、下側にコーパスに付属するキーワードのみを、それぞれランダムに 1 千件選びプロットした結果を示す。図 1 は縦軸を df_2/df 、横軸を df/N とした両対数グラフである。

4. 任意に切り出した部分文字列の存在範囲

図 1 の結果から、任意に切り出した部分文字列の反復度にはおおその上限と下限が存在する。(i)上限はおおよそ 0.6 ~ 0.8 であり、 df/N や、コーパスによって異なり、(ii)下限には、ポアソン分布を用いて文書に語が n 回出現することを考え、文書に語が 1 回以上出現する条件で 2 回以上出現する推定確率を考えた場合の値に近い。ある語がコーパス全体に含まれる数を cf とすると、ポアソン分布による推定確率は以下のようになる。

$$p_{\text{poisson}}(n) = 1 - \sum_{i=0}^{i=n} \frac{\lambda^i}{i!} e^{-\lambda} \quad \text{ただし} \quad \lambda = \frac{cf}{N}$$

$$p_{\text{poisson}}(2|1) = p_{\text{poisson}}(2) / p_{\text{poisson}}(1)$$

任意に切り出した部分文字列はほとんどの場合、(i)と(ii)を満足するおおよその範囲においてもれなく出現する。 df_2/df がポアソン分布による推定確率を下回る場合が存在するが、以下に示す特別な文字列であるため、存在範囲に含めなかった。

- 1 文字もしくは片言のひらがな、カタカナ、記号、一部の漢字。
- コーパスが扱う領域において非常に一般的な語。たとえば、論文抄録における「情報」や「システム」、新聞記事における「日本」などである。
- df_2 や df が共に高い常套句。たとえば、「について」「ている」などである。
- df_2 の値が df の値に対して非常に小さい常套句。たとえば、文書で 1 回もしくは非常に少ない回数しか出現しないもの。論文抄録における「評価に関する研究」「シミュレーションの結果について報告する。」「明らかにするため」などである。

5. 反復度の平均と頻度の関係

図 1 の結果より、任意に切り出した部分文字列の出現する領域において、キーワードは高い反復度を持ち、文書に繰り返し出現することがわかった。また、 df/N との関係調べるために、図 1 の結果に関して df/N 毎にグループ化して平均を取り、平滑化した結果を図 2 に示す。上側に任意に切り出した部分文字列を、下側にコーパスに付属するキーワードのみをそれぞれ 1 万件プロットした結果を示す。図 2 は縦軸を df_2/df 、横軸を df/N とした両対数グラフである。

図 2 の結果から、キーワードの反復度は、任意に切り出した部分文字列の反復度に比べ df/N に対して無相関であり、文献[1]での英語による分析結果に等しい。ただし、文献[1]の報告では df/N 毎のグループ化ではなく、ヘルドアウト推定を用い平滑化している。

任意に切り出した部分文字列は df/N が 10^3 倍変化する毎に df_2/df は 10 倍変化するのに対して、キーワードは df/N が 10^3 倍の変化する毎に df_2/df は 0.5 変化する。したがって、 df/N に対して df_2/df はほぼ無相関である。分析したコーパスに含まれる一文書は論文抄録でおおよそ 100 ~ 400 文字程度[6]、新聞記事はおおよそ 100 ~ 3000 文字程度であり、この条件では論文抄録においてコーパスに付属するキーワードの反復度はおおよそ 0.3 ~ 0.5、新聞記事においてコーパスに付属するキーワードの反復度はおおよそ 0.2 ~ 0.4 となった。

6. キーワード境界における反復度

前節での分析を基にキーワード境界における反復

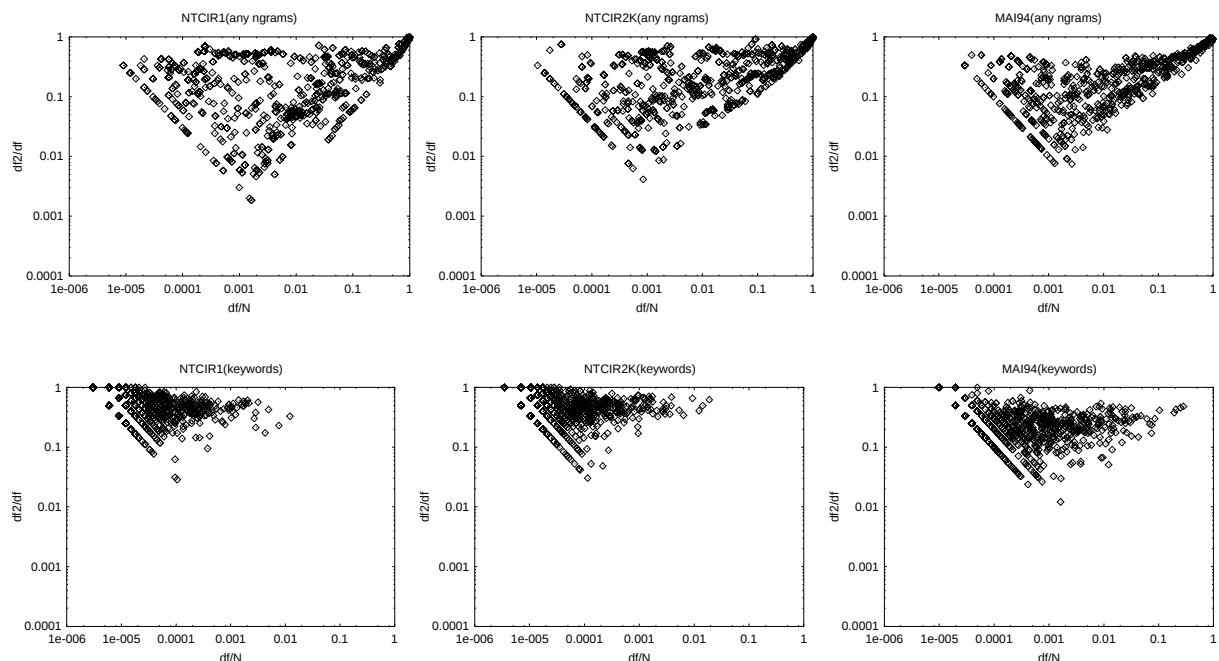


図 1 キーワードの分布と任意に切り出した部分文字列の分布

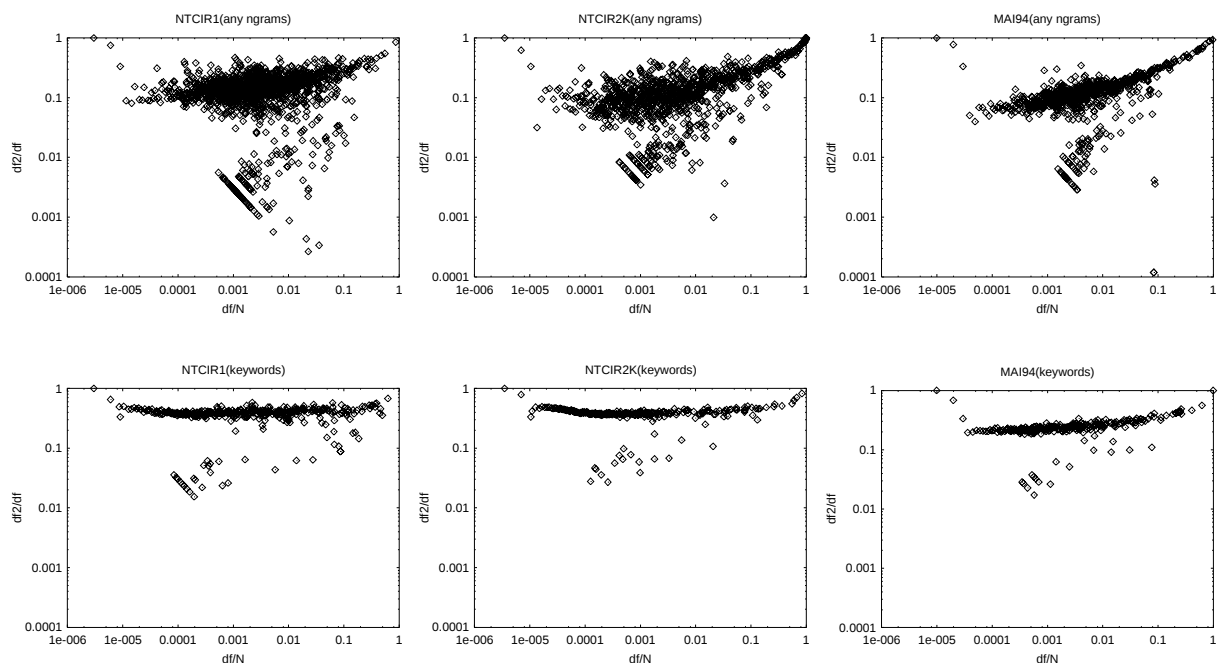


図 2 平滑化した反復度の分布

度の変化を考える．自立語のみで構成されたものは語であるが，自立語に付属語を加えて構成されたものは語ではなく句であると本論文では考えた．すなわち，ここで想定する語は「ひとまとまりの意味概念を表す」ものとする[7]．本論文では，反復度が自立語と付属語の境界を特定するために有用な特徴量ではないかと考え，キーワード（これは自立語に属するものである）

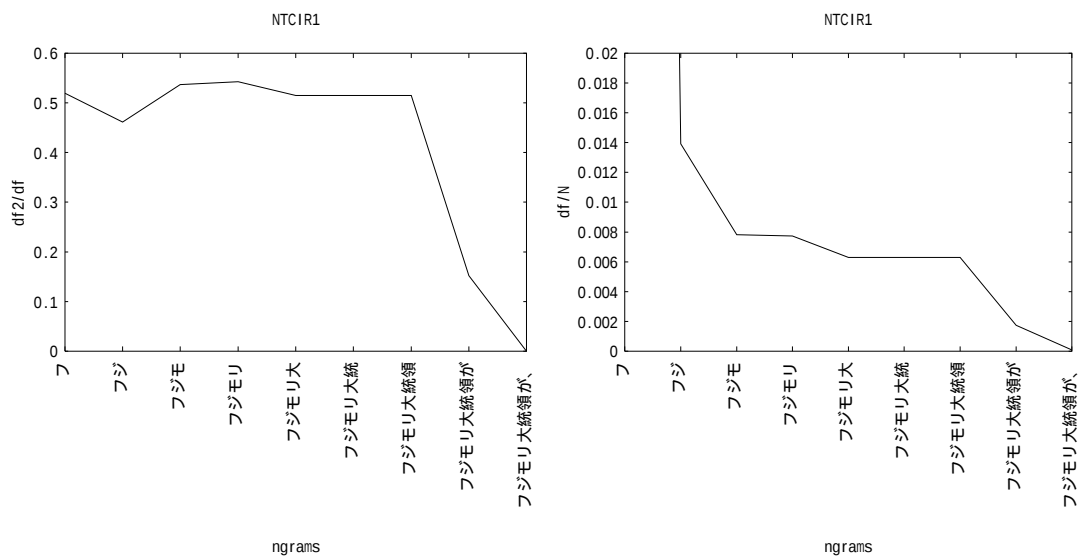
に，付属語である助詞を付加したときに反復度がどのように変化するかを分析した．

7. 反復度の推移

文字を最小要素とした文字列について，それぞれの文字を最小要素とした文字列について，それぞれの文字の接続関係について反復度を用いて考える！フジモリ大統領が「」という文字列に関して任意に切り出した

表 2 頻度調査「フジモリ大統領が」(NTCIR1)

cf	df	df_2	df / N	df_2 / df	部分文字列
84769	35313	18343	0.2947	0.5194	フ
3383	1668	770	0.0139	0.4616	フジ
2067	937	503	0.0078	0.5368	フジモ
2057	927	503	0.0077	0.5426	フジモリ
1533	756	389	0.0063	0.5146	フジモリ大
1533	756	389	0.0063	0.5146	フジモリ大統領
1533	756	389	0.0063	0.5146	フジモリ大統領
256	210	32	0.0018	0.1524	フジモリ大統領が
12	12	0	0.0001	0.0000	フジモリ大統領が,
92533	37559	19747	0.3134	0.5258	ジ
2089	951	509	0.0079	0.5352	ジモ
2057	927	503	0.0077	0.5426	ジモリ
1533	756	389	0.0063	0.5146	ジモリ大

図 3 df_2/df と df/N の値の推移

部分文字列の長さを、「フ」、「フジ」、「フジモ」と 1 文字ずつ増やし、頻度をそれぞれ計数した結果を表 2 に示す。

文献[1]によれば、英語において自立語は高い反復度を持ち、付属語は低い反復度を持つことが報告されている。日本語における分析においてその結果を目視すると、このような結果は同様であるが、さらに日本語における自立語の部分文字列は、意味のある語になるとは限らないが多くの場合高い値を持ち、自立語に助詞を追加した文字列の反復度は多くの場合それより小さい値を持つということがわかった。典型的な例を図 3 に示す。図 3 の左側には df_2/df の値の変化を示し、右側には df/N の値の変化を示した。

8. キーワードに助詞を加えた場合の反復度の変化

自立語の部分文字列は大きく分類すると、 $df/N > 0.5$ となる文字列と続く文字列が十分に予測

可能な文字列に分けることができる。 $df/N > 0.5$ となる文字列はコーパス全体において、ドキュメントを問わず頻出する文字列であり、この文字列の df , df_2 は共に大きい。それに対し、続く文字列が十分に予測可能な文字列とは次のようなものである。語を構成する文字の連続を考えた場合、文字列の文字数が増えていく毎にその文字列を含む用語の数は減る。多くの場合、語を構成する前の文字列において df_2/df の値には大きな変化が現れない。たとえば、「フジモリ大」という文字列に対して、続く文字列が「統領」であろうことは十分に予測可能である。表 2 からわかる通り、「フジモリ大」と「フジモリ大統領」の cf の値は等しい。このような文字列はキーワードと同様に文書に繰り返し出現する。逆に、キーワードに続く助詞を加えた文字列では、続く助詞が多岐に渡るため予測不可能である。キーワードに続いて特定の助詞が出現する頻度はキーワードが出現する頻度に比べ極めて少ない。

コーパスに付属するキーワードとそのキーワード

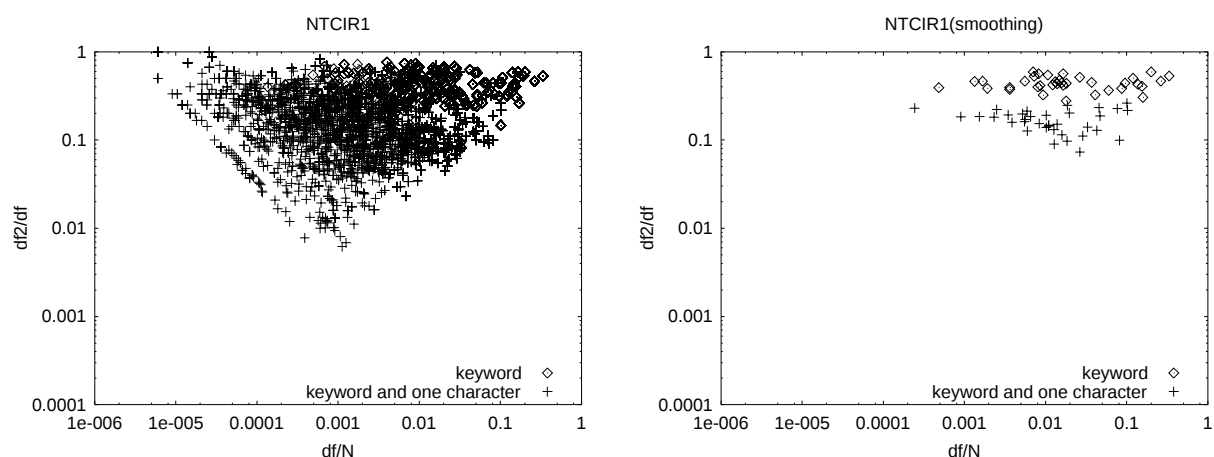


図 4 キーワードとキーワードに 1 文字を追加した文字列に関する反復度

に助詞を加えた文字列の反復度についてそれぞれ 4967 件を図 4 の左側に示す。キーワードに助詞を加えた文字列は実際にコーパスに現れた文字列である。また、図 4 の左側の結果を図 2 と同様に df/N でグループ化し平均を取ったグラフを図 4 の右側に示す。

ここで、キーワードに加えた助詞は語ではなく、コーパス中でキーワードに続いて現れた 1 文字である、追加する文字数を変化させて計数した結果、通常 1 文字で極端に反復度が減少し、続く文字を追加しても変化は小さかったためである。図 4 の結果から、分散は大きい、コーパスに付属するキーワードとそのキーワードに 1 文字加えた文字列の反復度は、平均を取ると明確な差が存在することがわかる。これによって反復度がキーワード境界を特定できる性質を持つことがわかる。

頻度によらないという反復度の持つ特徴はキーワード境界の特定には優れた性質である。一般的に語の持つ文書頻度はキーワードらしさとして有効であるとされているが、キーワード境界特定のための尺度として df/N を用いた場合(図 3 の右側を参照)、その値の変化は語の種類に従わず変化するためグラフでの変化は階段状になり、正しく境界を特定することは難しくなる。図 3 の右側の場合、部分文字列を「フジ」から「フジモ」と変化させた場合と、「フジモリ大統領」と「フジモリ大統領が」と変化させた場合に近い変化量を持っているが、キーワード境界に対応した変化ではないことがわかる。

まとめとして、反復度は、自立語の終端と付属語との境界に敏感に反応する。言い換えれば、文字を最小要素とした文章中において、自立語の境界を特定するための特徴量としても有用であることが観測できた。

9. まとめ

本論文では、反復度の英単語における分析結果に対し、日本語においてより一般的な状況を考え、その上で英単語における報告と同様に、反復度が語彙に強く依存するという分析結果を得た。

また、任意に切り出したすべての部分文字列の分析結果より、反復度がキーワード境界を特定するために有用であることを確かめた。

謝辞

本研究は平成 12 年度 IPA 未踏ソフトウェア創造事業のプロジェクトの一部であり、住友電気工業株式会社の援助による成果である。

参考文献

- [1] Kenneth W. Church: "Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2 ", *Coling*, pp.173--179(2000).
- [2] 武田善行, 梅村恭司: "キーワード抽出を実現する文書頻度分析", 計量国語学, Vol. 23, No.2, pp.65--90(2001).
- [3] Kageura, K. and Umio, B.: "Methods of Automatic Term Recognition: A Review", *Terminology*, vol.3, no.2, pp.259--289(1996).
- [4] Noriko Kando et al.: "NTCIR: NACSIS Test Collection Project", *20th Annual Colloquium of BCSIRSG*, Autrans, France, March 25--27(2001).
- [5] 毎日新聞新聞社: 毎日新聞データ, 91, 92, 93, 94, 95, 96, 97 年版.
- [6] Noriko Kando: "Overview of the Japanese and English IR Tasks at the Second NTCIR Workshop(Draft)", *Proceedings of the Second NTCIR Workshop Meeting*, pp.4-37--4-60(2001).
- [7] 松本祐治, 影山太郎, 永田昌明, 齋藤洋典, 徳永健伸: "岩波講座 言語の科学 3 単語と辞書", 岩波書店(1997).