文書長に依存しない文書定数

木村大翼[1)]　田中久美子[2)]
1) 東京大学工学部計数工学科　2) 東京大学大学院情報理工学系研究科

1　はじめに

二十世紀前半より文書の複雑さを表す特徴量として様々な指標が研究されてきており、今日に至るまで様々な言語モデルに基づいた多くの指標が提案されてきた。文長は各文書ごとに異なるが、文書の複雑さは文書ごとにある程度同じであると考えられるので、求める指標は文長によらず一定となることが望まれる。しかし、提案されてきた言語モデルはあくまで現実の文書を近似したものに過ぎず、モデル上で一定となる指標であることが数学的に示されていても、実際の文書ではなかなか一定にならないのが実情である。

このような指標は、古くは文書の著者判別問題を解く目的で追求されてきた [3]。この問題に対しては、昨今では機械学習などより高性能な方法が考えられる。とはいえ、文長に依存しない特徴量自体は、自然言語の複雑さを定量化する点で、興味深い側面を含んでいる。例えば、指標を用いて自然言語とプログラミング言語の複雑さの差や文書の難易度を計量することが考えられる。

このような指標については、Tweedie と Baayen がまとまった研究を行い、実際に 12 個の指標について、複数の英語の文書上で一定となるかどうかとそれらの判別力について報告している [5]。結論とすると、指標の中で文書の長さによらず一定であるのは 12 個のうち K(Yule, 1944)[6] と Z(Orlov,1983)[1] のみであるとのことである。だが、Tweedie らが扱った言語データは英語のみの 1M に満たない小規模なものであった。本稿では、この既存研究でよい結果となった K,Z 以外に、同じ観点から捉えられる二つの指標について、 英語以外の言語でも実験を行い、また、数百 M 程度の大規模データを扱い、既存の指標に対して値の一定性を調べることを目的とする。

2　様々な指標

2.1　K

指標 K は文書の語彙の豊富さを示す指標として 1944 年に統計学者の Yule によって提案された。今、文書の総単語数を N、単語の種類を V とし、文書中に m 回出現する単語の種類を $V(m,N)$ とすると、K は

$$K = C\Big[-\frac{1}{N} + \sum_{m=1}^{m=N} V(m,N)\big(\frac{m}{N}\big)^2\Big] \quad (1)$$

で定義される。ここで C は K の値が小さくなりすぎないようにするための係数であり、Yule は $C = 10^4$ とした。また、Yule は文書の生成モデルにつぼモデルと呼ばれる文書中の単語はランダムに現れるものとしたモデルを仮定しており、そのモデルにおいて N が十分大きい時には、この K の期待値が一定となることを数学的に証明することができる。

K が語彙の豊富さを表すことを以下簡単に説明する。まず式 (1) において $(\frac{m}{N})$ は文書中 m 回出現した単語が現れる確率を表す。よって $(\frac{m}{N})^2$ はそのような単語が連続で現れる確率である。ここで同じ単語が連続で現れる確率が大きい場合は文書の語彙が乏しい場合、確率が小さい場合は語彙が豊富な場合と見なすことができる。式 (1) より前者の場合は K の値は大きくなり、後者の場合は K の値は小さくなることがわかる。このように K は同じ単語が連続で現れる確率に基づいた語彙の豊富さを表す指標である。

2.2　Z

複雑系に関連した指標として Z と r の 2 つの指標を紹介する。まず Z について説明する。

文書中に現れる各単語の出現頻度は Zipf の法則に従うということが経験的に知られている。Orlov は 1983 年に Zipf の法則を拡張して、総単語数が N である文書の単語の種類 V の期待値 $E[V]$ が一つのパラメータ Z を用いて

$$E[V] = \frac{Z}{\log(pZ)}\frac{N}{N-Z}\log\Big(\frac{N}{Z}\Big) \quad (2)$$

と表せることを示した。ここで p は文書中に最も多く現れる単語の相対頻度であり、Orlov は単語数 N によらない一定値であると仮定している。式 (2) において Z の値が大きくなるにつれて $E[V]$ の値が大きくなるので指標 Z は文書の語彙の豊富さを表す指標だと解釈できる。

実際に Z を求める際には式 (2) において $E[V] = V$ とおいて反復法を用いて数値的に解く [1]。

2.3 r

r は本研究で新たに試みた指標であり、単語のネットワーク構造に着目したものである。

まず文書中の各単語から構成される無向グラフ $\Omega_L = (W_L, E_L)$ について説明する。文書中の単語の種類を V とすると、W_L は $W_L = \{w_i\}$ $(i = 1, ..., V)$ で定義される各単語を頂点とする頂点集合である。また、E_L は $E_L = \{(w_i, w_j)\}$ で定義される単語間のつながりを表す枝集合であり、2 つの単語 w_i と w_j が連続して現れる場合に枝が存在する。

さて、このようにして得られるグラフの各頂点の次数分布に着目する。グラフにおいて頂点の次数が k である確率を $P(k)$ とおく。

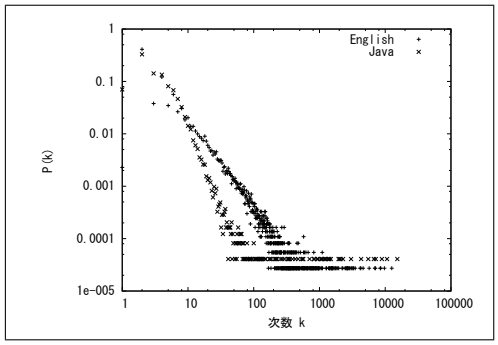

図 1　英語と Java の場合の次数分布

図 1 は英語と Java の場合の次数分布の両対数をとったグラフである。いずれもある次数まではほぼ直線になっている。このように単語のネットワーク構造の次数分布はベキ分布に従うことが知られている [4]。ベキ分布は

$$P(k) = ck^{-\gamma} \tag{3}$$

の形で表される。ここで c は規格化定数であり、$\sum_{k=1}^{\infty} P(k) = 1$ の条件から定まる。ここで式 (3) の両辺において対数をとれば

$$\log P(k) = -\gamma \log k + \log c \tag{4}$$

となりベキ分布は両対数グラフにおいて直線になることがわかる。

今、式 (4) の傾き $-\gamma$ に着目し、指標 r を

$$r = -\gamma \tag{5}$$

で定義する。

2.4 VM

これまでの指標は単語に着目したものであったが、最後に文字列に着目した指標 VM について紹介する。VM は近年 Golcher によって提案された指標であり [2]、接尾辞木の構造を利用したものである。

接尾辞木とは文字列の全ての接尾部を表した木構造であり、その枝には文字列が対応し、根から一つの葉に至る経路が一つの接尾部に対応している。

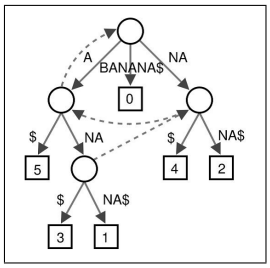

図 2　$BANANA$ の接尾辞木

図 2 は $BANANA$ という文字列に終端記号の\$を補った接尾辞木である。根から葉に至る経路で全ての接尾部である $BANANA$\$、$ANANA$\$、$NANA$\$、ANA\$、NA\$、A\$が現れている。

さて、この接尾辞木において根と葉以外のノードを内部ノードと呼ぶことにする。今、内部ノードの数を k、文字列の長さを T とおくと、VM は

$$VM = \frac{k}{T} \tag{6}$$

で定義される。

今、接尾辞木の構造から $0 \leq k \leq T-2$ であるので、式 (6) より $0 \leq VM < 1$ となる。Golcher の研究によると、興味深いことに自然言語、少なくともインド・ヨーロッパ語族においてはこの VM の値がおよそ 0.5 になるという実験結果が出ている。

3　実験結果

3.1　Tweedie らの研究の検証

Tweedie らは一人の著者によって書かれた英語文書を用いて K、Z の値が文書長によらず一定となることを示した。今回の実験では、他言語においても同様に一人の著者によって書かれた文書の各指標が一定となるかを確認するために日本語の「道標」、英語の *Two Years Before the Mast*、*The Sea Wolf*、フランス語の *Robert Burns Vol. I*、オランダ語の *Napoleon Geschetst Tweede omgewerkte druk*、スペイン語の *El Criterio* を用いて実験を行った。*The Sea Wolf* は Tweedie らが研究で用いた文書の一つである。

図 3 は各言語の K、図 4 は各言語の Z についての実験結果である。それぞれのグラフの横軸は文書の単語数の対数をとったものであり、縦軸はそれぞれ K、Z である。

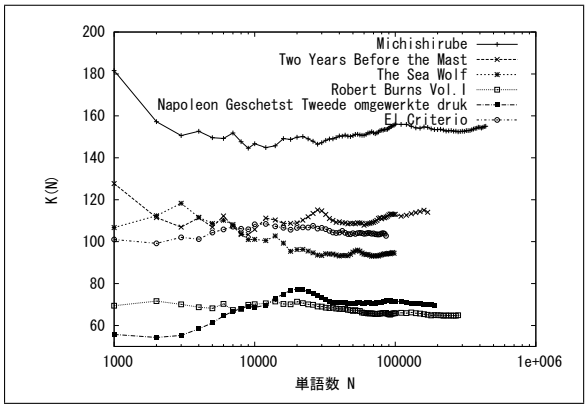

図 3　各言語の K (一人の著者による文書)

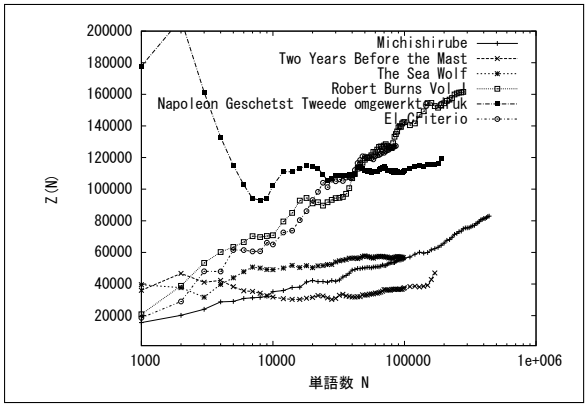

図 4　各言語の Z (一人の著者による文書)

まず英語文書の K,Z に注目すると Tweedie らの研究結果の通り、いずれの指標も文長によらず値がほぼ一定となっていることがわかる。さらに K についてはいずれの言語においても文長によらず値がほぼ一定となった。一方で Z については、単語数が一万を超えてからのオランダ語においてある程度一定となったが、それ以外の言語においては文長の log に対し値が増大する結果となった。よって Tweedie らの K、Z は文長によらず一定の値になるという結論は、K については正しいが、Z については英語、オランダ語以外の言語では正しくないということがわかった。

また同様の実験を r、VM で行ったところ、r の場合は値が一定とならなかったが、VM の場合は文字数が一万を超えたあたりからいずれの言語においてもおよそ 0.5 という一定の値となった。

以上から、一人の著者による文書の場合、言語、文長によらず値が一定となる指標は K、VM の 2 つであることが分かった。

3.2　大規模文書を用いた実験

様々な言語の大規模データを用いて実験を行い各指標の一定性を調べた。用いた言語データは日本語、英語、中国語、Java、Ruby であり、各データ量は自然言語でおよそ 100M、プログラミング言語でおよそ 50M である。一人の著者による数百 M の言語データは入手困難なので、新聞など複数の著者によって書かれた文書を用いた。また図 8 の各言語の VM については、日本語の場合は各文字をローマ字に、中国語の場合はピンインという中国語の読み方に変換して実験を行った。

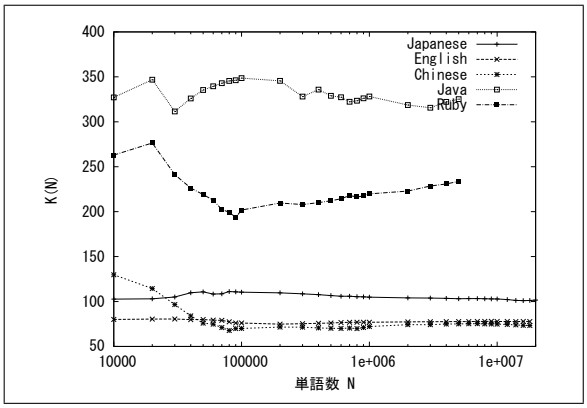

図 5　各言語の K (複数の著者による文書)

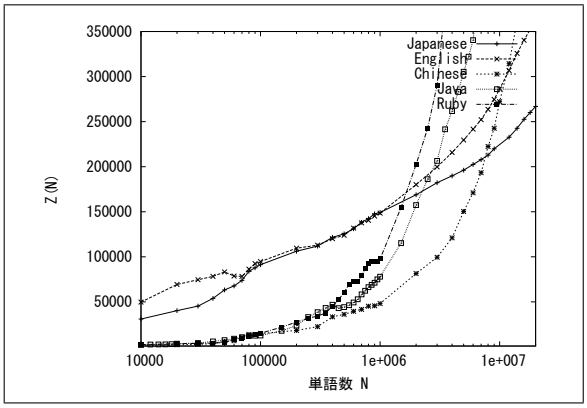

図 6　各言語の Z (複数の著者による文書)

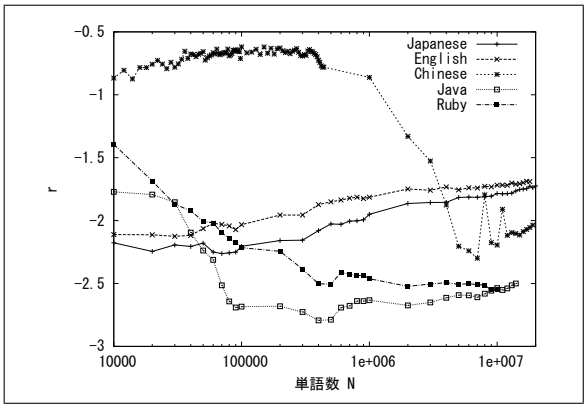

図 7　各言語の r (複数の著者による文書)

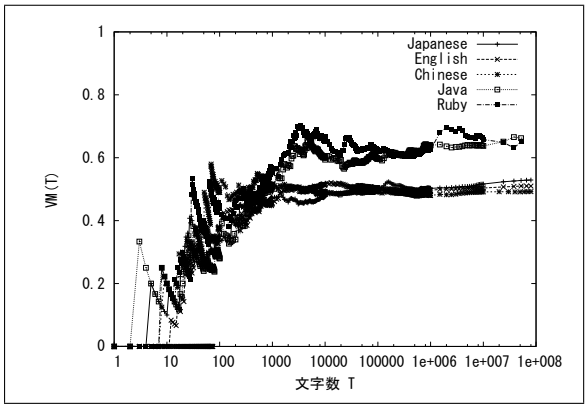

図 8　各言語の VM (複数の著者による文書)

図 5 は各言語の K、図 6 は各言語の Z、図 7 は各言語の r、図 8 は各言語の VM についての実験結果である。図 5、図 6、図 7 における横軸は文書の単語数の対数をとったものであり、縦軸はそれぞれ K、Z、r である。図 8 における横軸は文書の文字数の対数をとったものであり、縦軸は VM である。

まず K については先程よりも若干の値の変化が見られるものの、いずれの言語においても文長によらず値がほぼ一定となった。

次に Z と本研究で新たに試みた指標 r については、いずれの言語においても文長の対数に対して値が大きく変化してしまった。特に Z は単語数が百万を超えたあたりから変化がかなり大きくなった。

最後に VM についてであるが、自然言語においては日本語、中国語といった非インドヨーロッパ語族の言語であっても Golcher の研究結果と同様におよそ 0.5 の値をとるという結果が得られた。またプログラミング言語においては自然言語よりも若干の変化が見られるものの、およそ 0.65 というほぼ一定の値をとった。

4　考察

今回の実験から言語の種類、文書の長さによらず一定となる指標は K, VM の 2 つの指標であることがわかった。特に K については文書のランダム性が仮定されているにもかかわらず、文法制約などでその仮定が崩れている実際の文書においても値がほぼ一定となったということが興味深い。また、この 2 つの指標において自然言語とプログラミング言語の値に有意な差が見られた。これはこれらの言語間に複雑さの差があるためであると考えられる。

Z については一人の著者によって書かれた英語の文書においては値が一定となったが、複数の著者によって書かれた大規模文書では一定とならなかった。また英語以外の文書では値は一定とならなかった。なぜ英語の文書のみで一定となるのか、英語の文書において一定とならないのは複数の著者に原因があるのかそれとも文書が大規模であることに原因があるのかといった問題はこれからの課題である。

次に r については、日本語、英語の場合は文書が長くなると数十〜数百の次数をもつ頂点の数が増え、直線の傾きが次第に緩やかになってしまい一定とはならなかった。

最後に VM についてであるが、今回日本語、中国語の場合にはアルファベットに表記を変換して実験を行った。それぞれの本来の表記であるひらがな、カタカナ、漢字は文字の種類がアルファベットの種類よりもはるかに多いので、本来の表記のままで実験を行えば VM の値は 0.5 よりも小さくなると考えられる。またその場合に値が一定となるのかを調査することは今後の課題である。

5　まとめ

今日に至るまで数多くの文書の複雑さを表す指標が提案されてきたが、言語、文長に依存せず一定の値となる指標は現在のところ K、VM の 2 つのみである。この 2 つの指標においては自然言語とプログラミング言語間において値に有意な差が見られる。また、Tweedie らの研究結果の中の同一著者による英語文書において Z が一定であるということは他の言語においては必ずしも成り立たないということがわかった。

参考文献

[1] R. H. Baayen: *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht, 2001.

[2] F. Golcher: A stable statistical constant specic for human language texts. *Recent Advances in Natural Language Processing*, Borovets, September 2007.

[3] G. Herdan: *Quantatative Linguistics*. Butterworth, London, 1964.

[4] 増田直紀, 今野紀雄:「複雑ネットワーク」とは何か. 講談社, 東京, 2006

[5] F. J. Tweedie and R. H. Baayen: How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, vol. 32 (1998), pp. 323–352.

[6] G. U. Yule: *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, 1944.