h 指数を用いたテキストの特徴分析

鈴木崇史 † 冨坂亮太 †† 内山清子 † 相澤彰子 †

国立情報学研究所 † 東京大学大学院情報理工学系研究科 ††

t_suzuki@nii.ac.jp

本研究では，テキストの特徴分析のための指標として，h' – *type ratio* を提案する．従来提案されてきたテキスト指標の多くは，個々の文書に対して，絶対的な値として与えられてきた．本研究では，ジャンル判別，レコメンデーションなど，テキスト指標の多くの応用において，文書集合における特定文書の相対的な特性を考慮することに意義があると考え，このような指標として，新たに，h' – *type ratio* を提案する．これは，近年，科学計量学で提案された h 指数を，テキストデータに対応させ，改良したものであり，文書集合の特性を考慮にいれ，かつ，h 指数の利点を継承するものである．情報処理学会の二種類の論文本文データを対象として，提案指標の特徴を検討する．

1　はじめに

テキストを特徴づける指標は，従来から数多く提案されてきた．もっとも単純なものとしては，テキストの延べ語数，異なり語数などがあり，また，テキストに出現する頻度 1 や頻度 2 の語に注目したもの，頻度スペクトラム全体を考慮したものなど様々なものがある [7, 8]．これらテキストを特徴づける指標は，ジャンル判別やレコメンデーションなど，多くの応用に利用されている [6, 9]．

これら，従来のテキスト指標は，多くの場合，個々の文書に対して，絶対的な値として与えられてきた．これらの指標においては，文書集合の特性，また，文書集合における，特定文書の相対的な特性は，考慮されることが少なかった．しかし，ジャンル判別やレコメンデーションなど，実際の多くの応用で問題とされているのは，このような相対的な特性であり，求められているのは，これを反映する指標である．

このような指標として，本研究では h' – *type ratio* を提案する．本指標は，近年，科学計量学で提案され [5]，さかんに研究が行われている [1–4]，h 指数[1] を基礎として，これを，テキストデータに対応させ，また，改変したものである．h 指数は，研究者の論文生産性を測るための指標であり，「その研究者が公刊した N 本の論文のうち，h 本の論文が少なくとも h 回の被引用数をもち，他の $N-h$ 本の論文は，被引用数が h 未満である，以上を満たす最大の h の値」と定義される [5]．この指標は，研究者の公刊論文数と被引用数を同時に考慮した指標であり，多くの類似指標に比べて優れた特性をもつ.[2] 本研究では，研究者＝文書，公刊論文数＝文書に含まれる異なり語数，被引用数＝他の文書における語の利用数，すなわち *DF* 値，と対応関係を定義し，文書集合の特性を考慮にいれ，かつ，h 指数の利点を生かしたテキスト指標を提案する．

2　提案指標

本研究で提案する，テキストにおける h 指数は以下の通りとなる．

> h 指数: その文書が含む語のうち，h 種類の語が少なくとも，h 文書で用いられており，他の $N-h$ 種類の語は，h 文書未満で用いられている，以上を満たすような最大の h の値

例えば，h 指数が 30 である文書は，文書集合中，30 以上の文書で用いられている語を少なくとも 30 種類含むことになる．ここで，異なり語数と *DF* 値は，公刊論文数と被引用数の関係に比べて，大きなスケール差が生じ得る．そこで，以下のように h を補正する．

$$h' = \lfloor h \times \frac{max(DF)}{V(N)} \rfloor$$

[1] 実際には，h 指数は，ベキ則に従う現象一般に定義できる [4]．

[2] 例えば，被引用数の総和に比べて，被引用数上位の少数の論文に強く影響されないという特性をもつ [5]．

ただし，$max(DF)$ は，当該文書に含まれる語の最大の DF 値，$V(N)$ は，当該文書の異なり語数である．これによって，異なり語数と DF 値のスケール差を補正することができる．従って，h' 指数は，以下のように定義される．

> h' 指数: その文書が含む語のうち，h' 種類の語が少なくとも，h' 文書で用いられており，他の $N - h'$ 種類の語は，h' 文書未満で用いられている，以上を満たすような最大の h' の値

さらに，このように求められた h' 指数は，文書の異なり語数に強く依存すると想定される．そこで，以下の $h' - type\ ratio$ を提案する．

$$h' - type\ ratio = \frac{h'}{V(N)}$$

これにより，異なり語数による影響をより少なくし，かつ，h' の特徴を継承することができる．

3 実験の概要

情報処理学会の二種類の論文本文データを対象として，提案指標の特徴を検討する．データとして，情報処理学会論文誌ジャーナル (JNL)，情報処理学会誌「情報処理」(MAG) を利用する．JNL，MAG を併合した文書集合に対して，(a) 名詞全てを抽出したもの，(b) キーワードを抽出したもの，二種類のデータセットを作成し，提案指標の値を計算する．形態素解析には，MeCab (mecab.sourceforge.net) を利用し，名詞の抽出は，MeCab の品詞タグにもとづいて行う．キーワード抽出には，情報処理ハンドブックインデックスリスト [10] (5958 語) を利用し，テキスト，キーワードともに，形態素解析を適用した後，マッチングすることで，これを行う．(a) については，情報処理学会提供，JNL，2735 文書，MAG，1212 文書，(b) については，キーワードを 1 つ以上含む，JNL，2452 文書，MAG，1194 文書を対象とする．文書集合中に出現する名詞の $V(N)$ は 202586，キーワードの $V(N)$ は 3405 である．

$V(N)$ と h' の相関を確認した後，$h' - type\ ratio$ を計算する．提案指標の分布傾向に，顕著な特徴が見られるか否かを検討する．

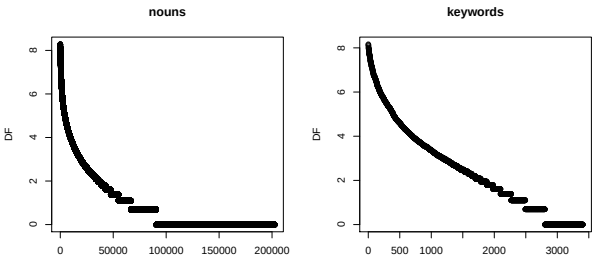

図 1　DF 値

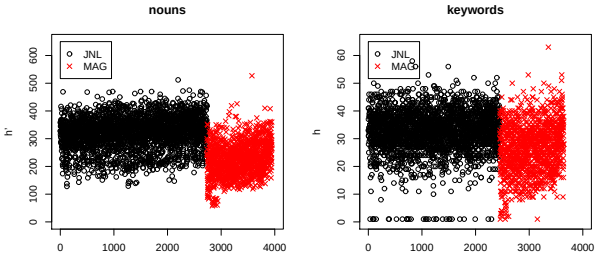

図 2　h'

表 1　異なり語数 ($V(N)$)，h' の中央値

	名詞		キーワード	
	$V(N)$	h'	$V(N)$	h'
JNL	1053.0	327	91.0	33
MAG	707.5	228	75.0	27

4 結果と考察

本節では，実験の結果と考察を示す．まず，4.1 節で，DF，$V(N)$，h' についての基礎的な結果を示し，$V(N)$ と h' の相関を確認する．続いて，4.2 節で，$h' - type\ ratio$ に関する結果を示す．

4.1 基礎的観察

図 1 は，名詞全てを用いた実験，キーワードを用いた実験，それぞれについて，個々の語の DF 値の対数をとり，降順に示したものである．図 2 は，同様に，二種類の実験について，個々の文書の h' 指数の値を示したものである．[3] 図 3 は，個々の文書の $V(N)$ と h' の関係を示したものであり，表 1 は，$V(N)$ と h' の中央値を示したものである．

図 3 左図から，名詞全てを用いた場合，(a)JNL が二群に分かれること，(b)MAG は，JNL 上位群

[3] 以下，JNL は○で，MAG は×で示されている．

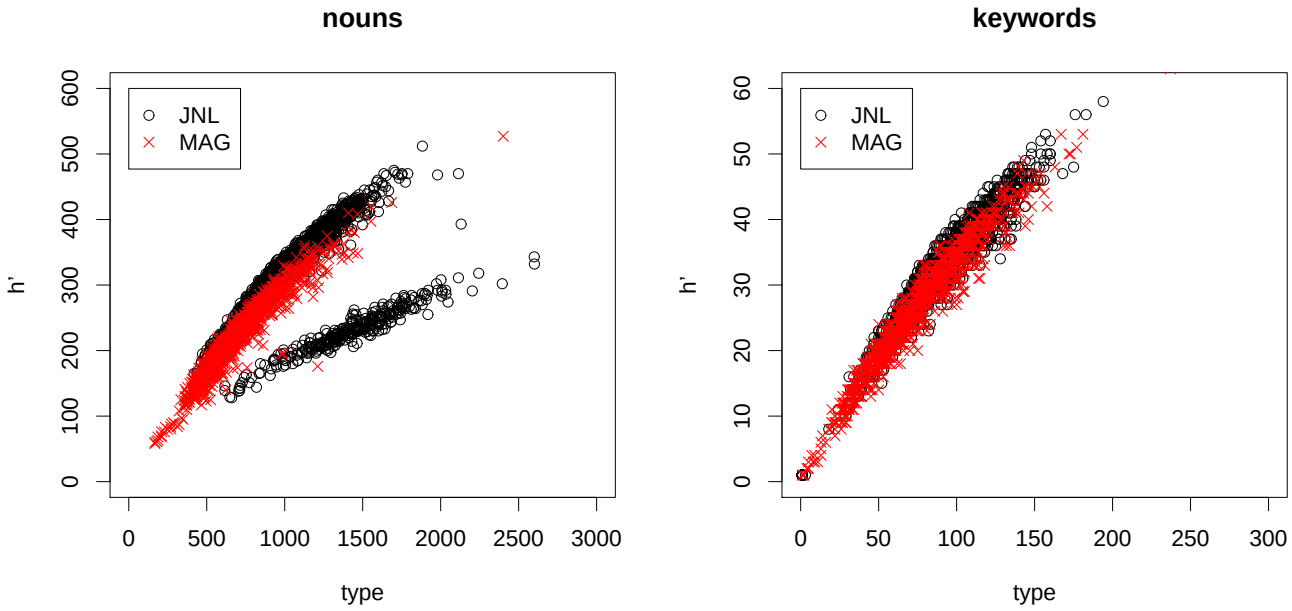

図 3　異なり語数 ($V(N)$) と h' の関係

より，やや低い値をとること，以上が確認される．図 3 右図から，キーワードを用いた場合，(a) の傾向は確認されず，(b) の傾向は，やや不鮮明ではあるが確認される．[4] また，図 3 から，$V(N)$ と h' は，高い正の相関をもつことが確認される．[5]

文書の特徴を質的に検討した限りでは，名詞全てを用いた実験が二群に分かれる理由は，今回用いたデータセットに，日本語論文と英語論文，双方が含まれているためであると指摘できる．[6]

4.2　h' – *type ratio*

図 4 は，個々の文書の $V(N)$ と h' – *type ratio* の関係を示したものである．4.1 節，図 3 で示した，h' と同様の結果が観察され，また，h' ほど顕著ではないものの，一定程度異なり語数への依存が観察される．

図 4 右図から，キーワードを抽出したデータに関しては，JNL と MAG の間に h' – *type ratio* の値に差があるか否か自明ではないため，ウィルコクソン順位和検定を適用し，この点を確認した．h' – *type ratio* が，$V(N)$ に依存することを考慮し，異なり数 0-50 語，50-100 語，100-150 語，150 語以上，これら四群にわけて検定を行った．

結果，全ての群に関して，$p < .01$ で，優位に JNL の方が MAG よりも h' – *type ratio* の値が高いことが示された．この結果は，JNL と MAG でキーワード使用の傾向が異なることを示している．

JNL の方が MAG より，h' – *type ratio* の値が高い理由は，今回用いたデータセットが，JNL2452 文書，MAG1194 文書であり，文書集合に対して JNL の方が比率が高く，JNL に特徴的な語の *DF* 値が高くなっていることによると指摘できる．

文書の特徴を質的に検討した限りでは，キーワードを抽出したデータについては，異なり語数が等しい場合，文書集合に対して支配的なトピック，すなわち，学会において支配的なトピック，あるいは，長い期間流行したトピックに関連する文書で h' – *type ratio* が高く，逆に，非支配的なトピック，すなわち，特殊なトピック，あるいは，一時的に流行したトピックに関連する文書では h' – *type ratio* が低くなっていると指摘できる．[7]

[4] この点は，4.2 節では，検定を行うことで，より厳密に確認する．

[5] より正確には，左図，すなわち，名詞全てを利用した場合には，(a) で確認された二群を区別した場合，高い正の相関が確認されることになる．

[6] 英語論文には，キーワードが出現しないため，キーワードを用いた実験では，このような傾向は見られない．

[7] 例えば，$V(N) = 87$(中央値) をもつ論文では，上位に「モバイルコンピューティング」，下位に，「自然言語処理」，「コンピュータと教育」，「人文科学とコンピュータ」などの分野の論文が含まれている．この点に関する厳密な検討は，今後の課題とする．

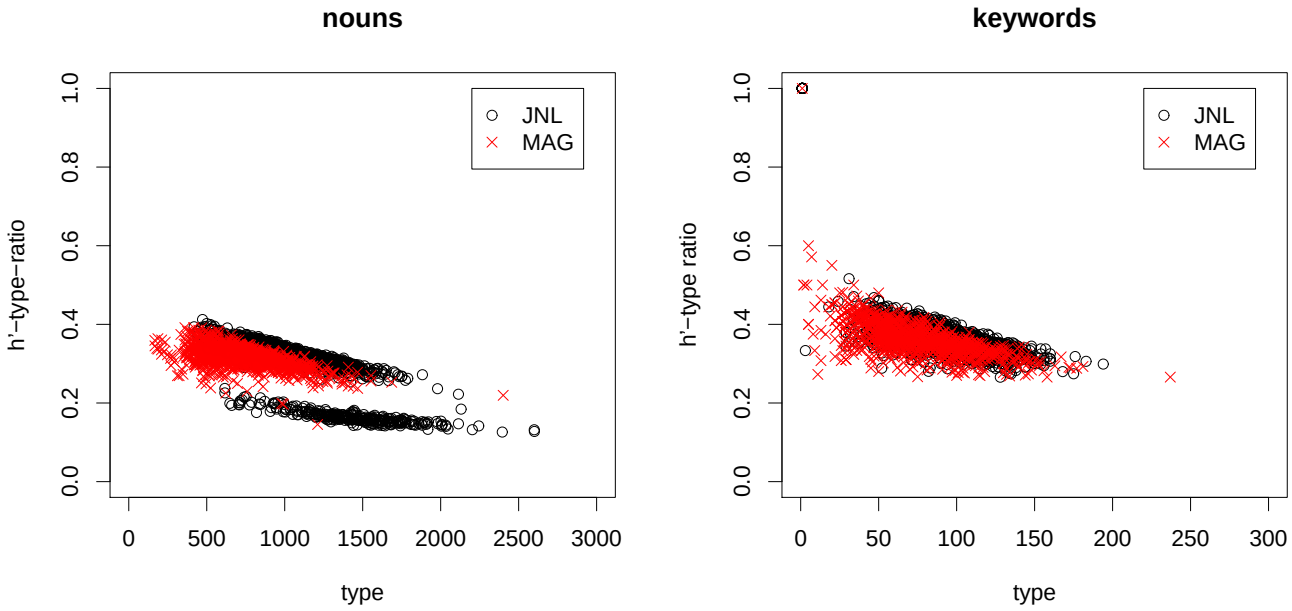

図 4　異なり語数 ($V(N)$) と h' – *type ratio* の関係

5　おわりに

本研究では，テキストの特徴分析のために，h' – *type ratio* を提案した．情報処理学会の二種類の論文本文データを対象として，提案指標の特徴を検討した．結果，(a) 名詞全てを用いた実験で，h' – *type ratio* の値によって，JNL が二群にわかれること，(b) 名詞全てを用いた実験，キーワードを用いた実験，それぞれにおいて，h' – *type ratio* の値が，二種類の論文本文データによって異なること，以上が示された．

提案した指標は，未だ，異なり語数に依存する指標である．今後，モンテカルロシミュレーションで異なり語数を統制し，指標の変化を検討することとしたい．また，文書集合におけるカテゴリー，すなわち，今回の実験例では，JNL と MAG の比率を変化させることで，指標がどのように変化するかを検討していきたい．

謝辞

情報処理学会刊行誌掲載論文本文データに関して，研究利用することを許諾していただいた，社団法人情報処理学会に感謝いたします．

参考文献

1) John Antonakis and Rafael Lalive. Quantifying scholarly impact: IQp versus the Hirsch *h*. *Journal of the American Society for Information Science and Technology*, Vol. 59, No. 6, pp. 956–969, 2008.

2) Rodrigo Costas and María Bordons. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, Vol. 1, No. 3, pp. 193–203, 2007.

3) Leo Egghe, Liming Liang, and Ronald Rousseau. A relation between h-index and impact factor in the power-law model. *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 11, pp. 2362–2365, 2009.

4) Leo Egghe and Ronald Rousseau. An informetric model for the Hirsch-index. *Scientometrics*, Vol. 69, No. 1, pp. 121–129, 2006.

5) Jorge E. Hirsch. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Science of the United States of America*, Vol. 102, 2005.

6) Anthony Kenny. *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities*. Pergamon Press, Oxford, 1982.

7) Fiona J. Tweedie and R. Harald Baayen. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, Vol. 32, pp. 323–352, 1998.

8) 影浦峡. 計量情報学. 丸善, 東京, 2000.

9) 金明哲, 村上征勝. 文章の統計分析とは. 甘利俊一, 竹内啓, 竹村彰通, 伊庭幸人（編）, 言語と心理の統計: ことばと行動の確率モデルによる分析, pp. 3–57. 岩波書店, 東京, 2003.

10) 情報処理学会（編）. 新版 情報処理ハンドブック. オーム社, 東京, 1995.