5U-08

用語の話題特定力を測る指標のための新パラダイム
－ ベースライン法の提案 －

久光 徹† 丹羽 芳樹† 辻井 潤一‡
†日立製作所 中央研究所 ‡東京大学大学院理学系研究科

1. はじめに

文書集合中の「話題を代表する傾向の強い」(= representative な)単語を選択・提示することは，文献検索インタフェイスや用語抽出において有効な技術である[1][2]。このような単語を捉えるための指標はこれまでも数多く提案されてきたが(review[3]及び[4]の第 2 節を参照)，それらには，高頻度で，かつ話題を担わない「不要語」を選択的に除去する能力の低さ，語の要・不要を判定するための閾値設定の困難，頻度の著しく異なる語間での指標値の比較が困難，等の問題があった。

これに対し，我々は，これらの問題を含まない指標として，

「全文書 D_0の単語分布 P_0と，ターム Tを含む文書すべての集合 $D(T)$に含まれる単語の分布 $P_{D(T)}$との距離 $Dist\{P_{D(T)}, P_0\}$を $Dist\{P_D, P_0\}$と比較して"正規化"した値」

で，T の representativeness(= $Rep(T)$)を計測する方法を提案した[4][5]。ここで D は，$D(T)$と同じ数の単語を含むランダムサンプリングされた文書集合である。前提としている，タームの representativeness に関する基本的な作業仮説は，

「representativeness な単語は，それを含む文書集合に何らかの特徴がある」

というものである。ここで，文書を特徴付ける指標の値は，一般には文書集合の大きさ自体に影響を受けるため，頻度の大きく異なるターム T_1, T_2 について，それぞれを含む文書集合の指標を比較する場合，文書集合の大きさから由来する影響を除外する必要がある。正規化はこのために必要であった。

上記の例では，「文書集合の特徴を測る指標」として，「文書集合中の単語の分布の，全文書の単語の分布との距離」を用いたが，基本の作業仮説に従う限り，元となる指標を他の指標に変更して，上記と同じ正規化手法で，新たな指標を生成することは自然であろう。すなわち，前報の手法は，より包括的な，representativeness 指標を定義するためのパラダイムとして拡張できるはずである。また，前報の指標自体についても，「それが単語に導入する順序は，他の指標を用いた場合とどのような関係があるのか」，「あるコーパス上で定めた指標が，異なるコーパスの上でどの程度有効なのか」など，前報では明らかにできなかった問題がある。

本報告の目的は，representativeness 指標を定義する新たな枠組み（これをベースライン法と呼ぶ）を提示し，これを用いて既報告の指標を含む新たな 3 指標を構成し，2 種類の古典的指標を含む諸指標の性質を，単語選別能力，指標の頑健性，単語の序列の間の順序相関の観点から明らかにすることである。

2. ベースライン法

前節で述べたように，本節では，単語の representativeness を測る為の指標を系統的に生成する方法を述べる。基本作業仮説は次のとおりである：

「あるターム Tが特徴的ならば，Tを含むすべての文書からなる集合 $D(T)$は，"平均的な"文書集合にくらべて何らかの特徴を持つ」

これに基づき，文書集合を特徴付ける何らかの指標を"正規化"して単語の representativeness を測る指標を生成する手続きは次のようになる：

「文書集合を特徴付ける指標 M に対し，$D(T)$に対する指標 M の値 $M(D(T))$と，$B_M(\#D(T))$とを比較し(例えば対数変換して比を取る)，新たな値を定義する」

ここで，$\#D(T)$は $D(T)$に含まれる単語の数をあらわし，$B_M(\bullet)$は**ベースライン関数**と呼ぶ実数値関数で，ランダムサンプリングされた文書集合 D に対し，$\#D$ から $M(D)$を推定する。こうして定義した指標を，M から**ベースライン法により生成された指標**と呼び，$Rep(\bullet, M)$と書くことにする。$M(\bullet)$を $Rep(\bullet, M)$の**原指標**と呼ぶことにする。

原指標 $M(\bullet)$として，$Dist\{P_{D(\bullet)}, P_0\}$を取るのが前報で示した指標であり，以下では単語分布間の距離を測るのに用いた対数尤度比(Log-likelihood ratio)にちなみ，$Dist\{P_{D(\bullet)}, P_0\}$を $LLR(\bullet)$と書き，LLR からベースライン法で生成された指標を $Rep(\bullet, LLR)$と書く。

$M(\bullet)$として考えうる他の興味深い原指標は，「$D(T)$における単語の異なり数」[6]，及び，「$D(T)$における単語分布のエントロピー」である。背景となる作業仮説は，

「あるターム Tが特徴的ならば，$D(T)$における単語のヴァラエティは相対的に少なくなる」

というものであり，基本作業仮説の特殊な場合である。そこで，$DIFFNUM(T)$ を $D(T)$における単語の異なり数，$ENT(T)$を $D(T)$における単語分布のエントロピーとして定義し，それぞれからベースライン法で生成された指標を $Rep(\bullet, DIFFNUM)$，$Rep(\bullet, ENT)$と書くことにする。

$B_M(\bullet)$は，全文書集合から様々な大きさの文書集合 D をランダムサンプリングして点集合 $\{(\#D, M(D))\}_D$ を得，これらを対数変換により $\{(\log(\#D), \log(M(D)))\}_D$ とした後，区分線形近似により求める。

$Rep(T, M)$を求めるとき，$\#D(T)$が大きい場合は，$D(T)$の替わりに，$D(T)$から適当な数の文書をランダムサンプリングした部分集合を用いて $Rep(T, M)$を求める。これにより計算量の低減だけでなく，ベースライン関数が安定して高精度に近似できる部分が利用できる等の利点がある。

ベースライン法で生成された representativeness 指標の特長は，(1)高頻度語と低頻度語の比較ができる。 (2) 任意の n に対し，単語 n グラムに適用できる。(3)閾値を系統的に設定できる，等である

3章 実験

$Rep(\bullet,LLR)$，$Rep(\bullet,DIFFNUM)$，$Rep(\bullet,ENT)$，*tf-idf*，単純頻度の五つの指標について，四つの実験を行った。

Ⅰ．日経新聞 1996 年分の通常の記事(人事記事などを含まない)158,000 記事中に，3 回以上出現した単語(約 86000 語)から 20,000 語を無作為抽出し，そのうちの 2,000 個を，検索内容の概観に現われることが「好ましい：A」「どちらでもよい」「好ましく

A Paradigm of Defining Term Representativeness Measures
– Baseline Method –
Toru Hisamitsu,† Yoshiki Niwa, † and Jun-ichi Tsujii‡
†Central Research Laboratory, Hitachi, Ltd.
‡Graduate School of Science, the University of Tokyo

ない：D」の3種類に分類し，上記20,000語をランダムに並べたときと，何等かの指標でソートしたときで，特定のクラスに分類された語が先頭から N 位までにいくつ出現するかという累積出現頻度により，指標の単語選別能力を比較する。

A と分類される語について，その結果を先頭5,000位まで示したグラフが図 1 であり，D と分類される語についての同様のグラフが図 2 である。

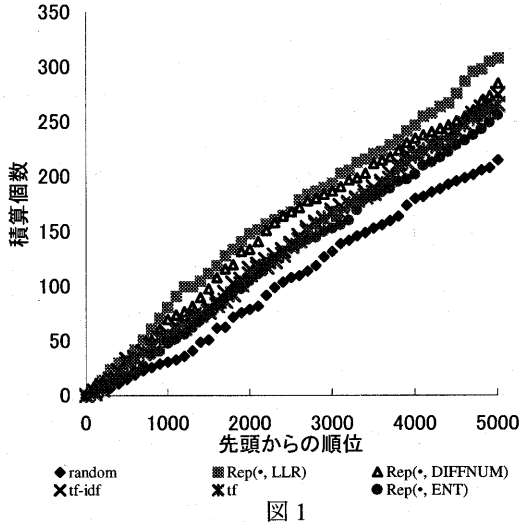

図 1
A と分類された語のソーティング結果

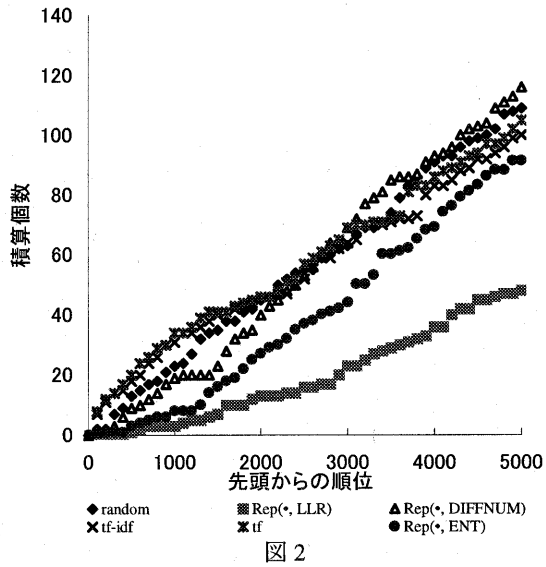

図 2
D と分類された語のソーティング結果

この範囲で，A に分類される語の優先順位を上げる力は，*Rep*(•,*LLR*)，*Rep*(•,*DIFFNUM*)，*Rep*(•,*ENT*)，tf-idf，頻度の順で強く，D に分類された語の優先順位を下げる力については，*Rep*(•,*LLR*)が圧倒的に強く，次いで *Rep*(•,*ENT*)で，他の指標はランダムと大差が無かった。

Ⅱ．*Rep*(•,*LLR*)を除く 4 指標が，上記 20,000 語からランダムに選んだ 2,000 語に対して導入する序列について，*Rep*(•,*LLR*)を基準とした順序相関を，Spearman，Kendall の順序相関係数により調べた結果，*Rep*(•,*LLR*)と他指標の相関は極めて低いことがわかった。これは，*Rep*(•,*LLR*)が，他の指標と異なる性質に注目していることを示している。

表 1 *Rep*(•, *DIFFNUM*)と他指標の順序相関

	Rep(•, *DIFFNUM*)	*Rep*(•, *ENT*)	*tf-idf*	*tf*
Spearman	-0.00792	-0.0663	0.202	0.198
Kendall	-0.0646	-0.0553	0.161	0.153

Ⅲ．*Rep*(•,*LLR*)について，I で用いた 158,000 記事(以下 NK96-ORG と称す)と別に，大きさ・分野の異なる 6 種類の文書集合を用意した。すなわち，日経 1996 年分からランダムに選んだ 5,0000 記事(NK96-50000)，100,000 記事(NK96-100000)，20,0000 記事(NK96-200000)，及び日経 1998 年分から選んだ 15,8000(NK98-158000)記事，NII-NACSIS テストコレクション[7]から選んだ 15,8000 個の日本語アブストラクト(NC-158000)，及び NACSIS テストコレクションの日本語アブストラクト全体(NC-ALL)である。

これらの 7 コーパス上でそれぞれ B_{LLR}(•)を生成し，NK96-ORG 以外のコーパスで生成した B_{LLR}(•)も NK96-ORG 上での正規化に用いることにより，NK96-ORG 上に本来の *Rep*(•,*LLR*)を含め 7 種類の指標を構成した。NK96-ORG 本来の *Rep*(•,*LLR*)が NK96-ORG 内の単語集合に導入する序列と，他の 6 種類が導入する序列の間の相関を，Spearman，Kendall の順序相関係数により調べたところ，NII-NACSIS 系コーパスで，それぞれ順序相関が 90%強，78 強%となったほかは，両相関係数ともほぼ 100%で，極めて高い順序相関があった。しかも，これらを用いて I と同様の実験を行い，図 1, 2 と同様の可視化を行ったところ，殆ど区別はみられなかった。

4. まとめ

タームの representativeness 指標を定義するための，ベースライン法と呼ぶ枠組みを提示した。これを用いて，「文書集合中の単語分布の全体の単語分布からの距離」，「文書集合中の単語異なり数」，「文書集合中の単語分布のエントロピー」をそれぞれベースライン法で正規化した 3 種の指標 *Rep*(•,*LLR*)，*Rep*(•,*DIFFNUM*)，*Rep*(•,*ENT*)を構成し，頻度,及び *tf-idf* と併せて，単語選別能力を調べた結果，*Rep*(•,*LLR*)は，他を圧倒する有用・不要単語の選別能力を持つことがわかった。また，*Rep*(•,*LLR*)が単語に導入する序列は，他の指標のそれとは順序相関が低い。さらに，*Rep*(•,*LLR*)のベースライン関数は，正規化能力においてコーパス依存性が低いことも確認した。これは，実用上重要な性質である。

さらに簡便な方法で(できれば解析的に)ベースライン関数を推定できることが望ましいため，現在，*Rep*(•,*LLR*)より鋭敏で，かつベースライン関数がより簡単に推定できる指標を開発中である。

参考文献

[1] Niwa, Y., Nishioka, S., Iwayama, M., Takano, A., and Nitta, Y. (1997). Topic graph generation for query navigation: Use of frequency classes for topic extraction. *Proc. of NLPRS'97*, pp. 95-100.

[2] Niwa, Y., Iwayama, M., Hisamitsu, T., Nishioka, S., Takano, A., Sakurai, H., and Imaichi, O. (2000) DualNAVI - dual view interface bridges dual query types, *Proc. of RIAO 2000*

[3] Kageura, K. and Umino, B. 1996. Method of automatic term recognition: A review, *Terminology* 3(2): 259-289.

[4] 久光徹, 丹羽芳樹, 辻井潤一 (1999). タームの representativeness を測るための一指標, 情報処理学会第 59 回全国大会論文誌, 2P-03.

[5] Hisamitsu, T., Niwa, Y., and Tsujii, J. (1999) Measuring Representativeness of Terms, *Proc of IRAL'99*, pp.83-90.

[6] 寺本陽彦, 宮原豊, 松本俊二 (1999). 類似文書検索のためのタームの共起語分布分析による計算, 情報処理学会第 59 回全国大会論文誌, IP-06.

[7] Kando, N., Kuriyama, K., and Nozue, T. (1999). NACSIS test collection workshop (NTCIR-1), *Proc. of the 22nd Annual International ACM SIGIR Conf. on Research and Development in IR*, pp.299-300.