単語の話題性尺度計算に関する調査研究

１ はじめに

大規模な文書集合の内容を要約するために，文書集合中の「話題を代表する傾向の強い(=representative な)用語(単語列)」を選択する技術は，テキストマイニングにおける基盤技術であり，辞書の自動構築や知識発見のために必須である。用語が何らかの話題を代表する力の強さ，すなわち「話題性」を示す量を representativeness と呼び，用語の representativeness を測る有効な尺度を開発することを目的に，我々はこれまでに Baseline 法と呼ぶ，用語の representativeness 尺度を構成するための基本方法を提案し，同法を用いていくつかの尺度を構成してきた[1][2][3][4][5][6][7]。本報告では，これまでに開発した representativeness 計算原理を整理した形で述べ，(2)で述べる，連想計算エンジン GETA を利用した representativeness 指標を計算するためのプログラムの利用について説明する。

２ これまで得られた知見のまとめ

2.1 従来の尺度

用語が何らかの話題を代表する力の強さを representativeness と呼ぶ。representativeness の概念は従来明示的に提唱されたことはないが，これと関連するものとして，情報検索の分野では「用語の重要度」に関する尺度が索引語の重み付けのために導入されてきた[8]. そこでは用語の文書内頻度，用語の全文書における出現頻度，注目する用語の分布の偏り等に基づき，さまざまな尺度が提案されている。なかでも最も簡単でかつ広く利用されているものに *tf-idf*[9]があり，あらかじめ文書がカテゴリに分類されている場合には，カテゴリごとの用語の出現頻度の偏りをχ^2 検定を利用して測る手法も提案されている[10]。自然言語処理の立場からは，相互情報量[11]，対数尤度比[12]など，「隣り合う語のまとまりの強さ」に注目してタームを選ぶ一連の方法と，ターム同士の包摂関係や、前後に出現する単語の豊かさ等に着目する尺度 C-value[13], NC-value[14], *Imp*[15]等が導入されている。しかし，従来の尺度には，

(1) *tf-idf* (もしくはその類似手法)等の古典的手法は，語の頻度の寄与が大きすぎる傾向があり，高頻度不用語の排除が困難である。

(2)語のカテゴリ間での分布の違いを比較する方法では，あらかじめ文書が分類されている必要があるが，この制約は強すぎる。

(3) 隣り合う単語の共起の強さを利用する手法では単独の単語の重要度が評価できない。また，本来この型の尺度と representativeness との関係は不明である。

(4)閾値の設定に理論的な根拠が与え難い。

等の問題があった。

2.2 Baseline 法の原理

我々は, このような問題の無い尺度を構成するための基本方式 Baseline 法を開発してきた。ここでは, これまでの研究[1]-[7]の結果を整理して概略を述べる。

Baseline 法に従って用語の representativeness を定義するための第一の仮説は以下のものである：

「用語 T が特徴的ならば, T と"共起する単語の集合"を $Co(T)$ として $Co(T)$は,
"平均的な"文書集合にくらべて何らかの特徴を持つ。」

これは, Firth が

"*You shall know a word by the company it keeps*. [16]"

と表現した, 広く受け入れられている考え方を換言したものである。ここで, "T と共起する単語の集合"の定義に応じて $Co(T)$はさまざまに定義しうる。

「単語集合に対する尺度」と呼ぶ, 単語集合から実数への写像を用いてこの仮説をさらに換言する。すなわち、単語集合に対する尺度 M を一つ固定し,「用語 T の特徴量」を, $M(Co(T))$により定義すると, 仮説は次のように換言できる:

「用語 T が特徴的ならば, $M(Co(T))$は, "平均的な値に比べて顕著な"値を持つ。」

従って、

(1) M の選択、
(2) $Co(T)$の定義,
(3) "平均的な値な値"の定義

が適切にできれば, T の特徴量がうまく定義できることになる。

しかし, これまでに考案されてきた単語集合上の尺度は, 一般に, 単語集合の大きさ(＝含まれる単語数)に依存して, その値が系統的に変動する[17](多くは単調増大または減少)性質があり, その結果、頻度が極端に大きい(または小さい)単語の特徴量が他のそれより大きいという結果となりがちであった[I]。しかし T の特徴量を単語集合上の尺度 M を介して定義することの本来の意図は, T の頻度とは独立した観点から用語の性質をとらえようとするものであるから, これは本来の目的を満たしているとはいい難い。

Baseline 法の原理は, 上に述べた,「$M(Co(T))$の値が $Co(T)$が含む単語数に依存して系統的に変動する」部分を推定し, それを $M(Co(T))$の値から分離することにより, $M(Co(T))$の値から,「T の頻度と独立な成分」を抽出することにある。これを M の「単語数に関する正規化」または単に「正規化」と呼ぶのであった。つまり, Baseline 法とは,「単語集合上の尺度を介して単語の特徴量を定義する枠組みにおける, 単語集合上の尺度の正規化方法」である。正規化された M を用いて用語 T の representativeness 尺度を定義すれば, 異なる頻度を持つ用語 Tの間で尺度の値を比較することが意味を持つ。さらに, このようにして定義した尺度には, 用語の representativeness の有無を決める閾値が合理的に設定できる等の特長があり, 用語の特徴量として優れた性質を持つ。

[I] 用語 T の頻度と, $Co(T)$の大きさには正の強い相関があるため, $Co(T)$の値が T の頻度に対して大きな単調増大または減少の傾向がある場合（多くの場合、実際にそうなのであるが）, T_1 と T_2 の特徴量 $M(Co(T_1))$と $M(Co(T_2))$の大小は, T_1, T_2 の頻度の大小だけで決まってしまうからである。

2.3 Baseline 法の適用手順

我々は，情報検索を意識した最も基本的な共起の定義として， $Co(T)$を"T を含む文書全体の単語の集合（重複を許す）" $D(T)$と定義する。$D(T)$は「文書の集合」を介して構成されるため，M は文書の集合上の関数とも見なせる。これを強調して，M を**文書集合に対する尺度**とも呼ぶ。この場合，$D(T)$は単語集合でなく，元の文書の集合と見なす。「$M(D(T))$の値が $D(T)$が含む単語数に依存して系統的に変動する」部分は，文書単位でランダムサンプリングして生成した文書集合 D_{rand} に対して，D_{rand} が含む単語数$\#D_{rand}$から $M(D_{rand})$を推定する関数 $B_M(\bullet)$を用いて， $B_M(\#D(T))$により与える[II]。M の値は，$M(D(T))$と $B_M(\#D(T))$の比を取るなどして正規化する。正規化された M を用いて定義した用語 T の representativeness 尺度を，**Baseline 法により M を正規化した尺度**又は単に ***M*を正規化した尺度**と呼び $Norm(M)$（正確には $Norm(M)(\bullet)$）と書くことにし， M を $Norm(M)$の**原尺度**と呼ぶ（[1]-(2)」では，より詳しく**話題性原尺度**と呼んでいる）。$B_M(\bullet)$を Baseline 関数と呼び，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得られる点集合 $\{(\#D,\ M(D))\}_D$ から近似により求める。$B_M(\bullet)$の求め方に関する詳細は 2.6 で述べる。

2.4 これまでに構成した尺度

文書集合に対する話題性原尺度 M を Baseline 法により正規化した尺度は，

○ 頻度が著しく異なる用語の間で尺度の値の比較ができる

○ representativeness の有無を決める閾値が合理的に設定できる

等の優れた特長を持つ。原尺度のうち，$D(T)$の単語分布を特徴付ける代表的なものとして，

◆ 用語 T を含む全ての文書集合 $D(T)$中の単語分布 $P_{D(T)}$と，全文書集合中の単語分布 P_0 の間の距離：$DIST(D(T))$

◆ $D(T)$における単語の異なり数：$DIFFNUM(D(T))$

◆ $D(T)$における単語分布のエントロピー：$ENT(D(T))$

等が考えられるが，中でも，$DIST$ を Baseline 法で正規化した尺度 $Norm(DIST)$は，既存の諸尺度や $Norm(DIFFNUM)$，$Norm(ENT)$等と比較して，特徴単語の選別能力が大きく優れ，用語の抽出実験においても有用性が確認された[5][6]。

2.5 *Norm*(*DIST*)の詳細

$DIST(D(T))$は定義により，用語 T を含む全ての文書集合 $D(T)$中の単語分布 $P_{D(T)}$と，全文書集合中の単語分布 P_0 の間の距離である。単語分布間の距離の計測の方法としては，

(1) 対数尤度比(log-likelihood ratio),

(2) Kullback-Leibler divergence,

(3) transition probability,

(4) vector-space model (cosign 法)

等が考えられるが，実験の結果対数尤度比を用いることにした．すなわち，単語出現の独立性の仮定のもとで，$D(T)$中の単語分布と全コーパス D_0 中の単語分布が独立であると仮定し

[II] ランダムサンプリングの基本単位は $Co(T)$を定義する共起単位に従う。今の場合は，1 文書である。

た場合の最尤推定値と，$D(T)$中の単語分布と全コーパス D_0 中の単語分布が同一であると仮定した場合の最尤推定値との比の対数値である。$\#D(T)$を$D(T)$に含まれる単語の延べ数，$\#D_0$を D_0に含まれる単語の延べ数，$\{w_1, \ldots, w_n\}$を D_0に含まれる全異なり単語，w_iの $D(T)$中での出現回数を k_i, D_0中で出現回数を K_iとしたとき，$DIST(D(T))$は以下で与えられる:

$$\sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(T)} - \sum_{i=1}^{n} k_i \log \frac{K_i}{\#D_0}.$$

2.6 ベースライン関数と正規化についての詳細

我々は尺度 M に関する Baseline 関数 $B_M(\bullet)$を得る際，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，これらを両側対数をとって$\{(\log(\#D), \log(M(D)))\}_D$に変換し，区分線形近似により得た[1]-[7]。

実際に $Norm(M)(D(T))$を求めるときは，$\#D(T)$が大きい場合，$D(T)$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 $Norm(M)(D(T))$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高精度に近似できる原点に近い部分が利用できる。

日経新聞1996年版を用いた実験によれば，原指標M=$\{DIST, DIFFNUM, ENT\}$において，1996年版の，様々なサイズ(2,000記事程度から300,000記事程度) の部分集合を全コーパスと見なしてベースライン関数を求めたところ， #D が1,000から20,000程度の間は，コーパスサイズに依らず，安定して良く近似できることがわかった。そこで，この区間において近似関数$B_M(\bullet)$を求め，$1000 \le \#D(T) \le 20{,}000$を満たす$T$について，値$Norm(M)(D(T))$を，$M(D(T))$に$B_M(\bullet)$による正規化を施した値により定義する。すなわち，

$$Norm(M(D(T))) = \frac{\log(M(D(T)))}{\log(B_M(\#D(T)))}.$$

ここで，「する」のように著しく$\#D(T)$が大きい単語の場合にも上記のベースライン関数の有効域を用いることを可能にするために，$20{,}000 < \#D(T)$となるような T に対しては，$D(T)$として150文書程度をランダム抽出し，$1000 \le \#D(T) \le 20{,}000$を満たすようにしてから$Norm(M)(D(T))$を計算した．これは同時に計算量の低減にもつながる。

ランダムサンプリングして生成した文書集合Dにおいては，$Norm(M)(D)$は，さまざまなコーパスにおいて平均 Avr が 1 に極めて近く(1±0.01)，標準偏差をδ(δの値も安定しており，0.05±0.001)として，$Norm(M)(D)$の値が $Avr \pm 4\delta$ を越えるものは1パーセント未満であった。よって，ターム T が *representative* であると判断するための閾値として，すべての M に対して $Avr + 4\delta$を設けることが妥当である。

図 1(a)は M=$DIST$ のとき，いくつかのターム T についての$\{(\#D(T), DIST(D(T)))\}_T$のプロットとベースライン曲線を示し，図 1(b)は，図 1(a)での$\{(\#D(T), DIST(D(T)))\}$の値を正規化した結果を示している。例えば正規化前だと，$DIST(D(\text{“する”}))$> $DIST(D(\text{“暗号”}))$であるが，正規化によりこれが逆転することが分かる。

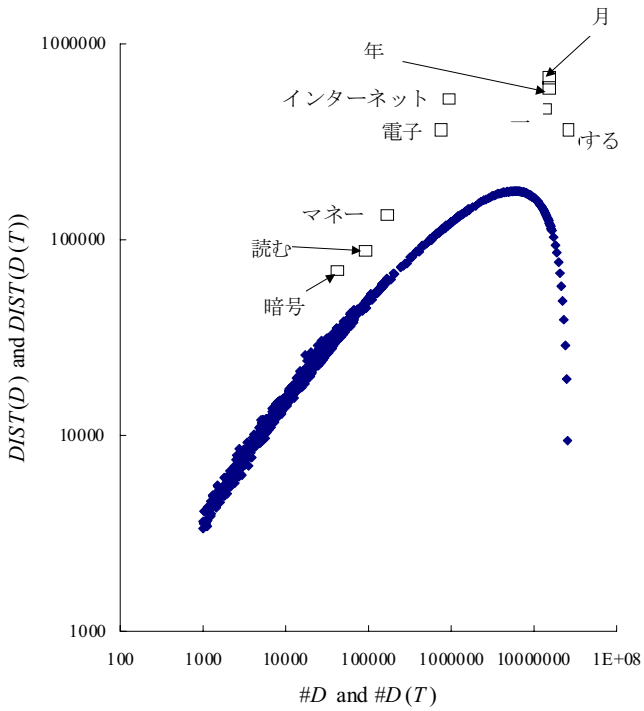

図 1(a)
*DIST*のベースライン曲線と, サンプルタームの値のプロット

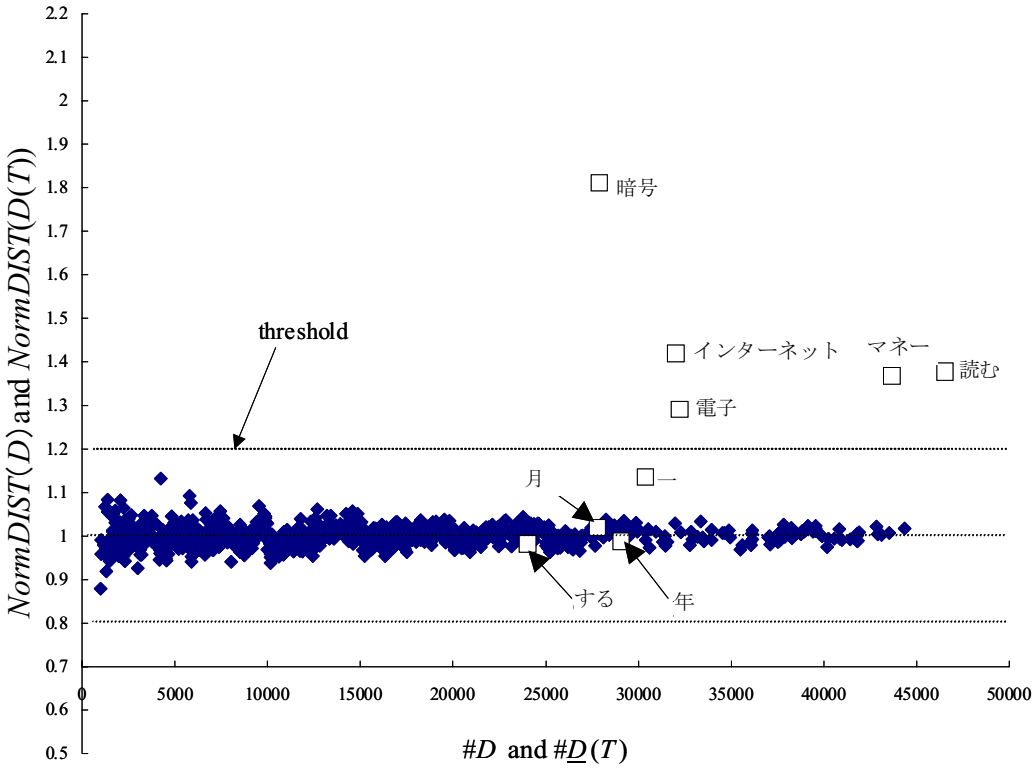

図 1(b)
正規化の効果

3 話題性原尺度計算プログラムの使い方

「文書集合に対する話題性原尺度計算モジュール」では5つの関数を解説している。そのうち dr_doc_repres_raw は, コーパス中の任意の文書集合 D に対して, 話題性原尺度の例として $DIST(D)$と $DIFFNUM(D)$を計算する関数である。入出力の型の詳細については[1]-(2)に譲り, ここではこの関数を応用してコーパス中の任意の単語 w について $Norm(DIST)(D(w))$と $Norm(DIFFNUM)(D(w))$を計算するための指針について述べる。

$Norm(DIST)(D(w))$を求めるには, $DIST$の Baseline 関数$B_{DIST}(\bullet)$を求めて正規化する。関数$B_{DIST}(\bullet)$を求めるためには, さまざまな数の文書をランダムサンプリングした文書集合 D に対して dr_doc_repres_raw で $DIST(D)$を求め, 対数線形近似を使うならば $\{(\log(\#D), \log(DIST(D)))\}_D$ をプロットして近似して求める。$DIST(D(w))$ は, $D(w)$ を求めて dr_doc_repres_raw に渡せば求まる。$Norm(DIST)(D(w))$ は $\log(DIST(D(w)))$ を $\log(B_{DIST}(\#D(w)))$ で割れば求められる(対数線形近似を使った場合)。$Norm(DIFFNUM)(D(w))$も全く同様にして求められる。

dr_doc_repres_raw 中で計算している原尺度は $DIST$ と $DIFFNUM$ のみであるが, 定義式を書き換えれば任意の原尺度 M を計算することができ, 従って $Norm(M)$を求めることができる。

参考文献

[1] Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M., and Takano, A. (1999a). Term Extraction Using A New Measure of Term Representativeness, in *Proc. of NTCIR'99*, pp475-481.

[2] Hisamitsu, T. Niwa, Y., and Jun-ichi Tsujii (1999b). Measuring Representativeness of Terms, in *Proc. of IRAL'99*, pp.83-90.

[3] 久光徹 丹羽芳樹 (1999a) タームの representativeness を測る,情報処理学会研究会報告(自然言語処理研究会), Vol. 99-NL-133, pp.115-122.

[4] 久光徹 丹羽芳樹 辻井潤一 (1999b) 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, pp.3-63-3-64.

[5] Hisamitsu, T., Niwa, Y., and Tsujii, J. (2000) A Method of Measuring Term Representativeness - Baseline Method Using Co-occurrence Distribution-, *Proceedings of COLING2000*, pp.320-326

[6] 久光徹 丹羽芳樹 辻井潤一 (2000). 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 ―, 情報処理学会第 61 回全国大会論文誌, pp.3-197-3-198.

[7] Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M. and Takano, A. (2001) Extracting Terms by a Combination of Term Frequency and a Measure of Term Representativeness, *Terminology* (to appear)

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

[9] Salton, G. and Yang, C. S. (1973). On the Specification of Term Values in Automatic Indexing. *Journal of Documentation* 29(4), pp.351-372.

[10] Nagao, M., Mizutani, M., and Ikeda, H. (1976). An Automated Method of the Extraction of Important Words from Japanese Scientific Documents, Trans. *of IPSJ*, 17(2), pp.110-117.

[11] Church, K. W., and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics* 6(1), pp.22-29.

[12] Dunning, T. (1993). Accurate Method for the Statistics of Surprise and Coincidence, *Computational Linguistics* 19(1), pp.61-74.

[13] Frantzi, K. T., and Ananiadou, S., and Tsujii, J. (1996). Extracting Terminological Expressions, 情報処理学会自然言語処理研究会報告, NL112-12, pp.83-88.

[14] Maynard, D. and Ananiadou, S. 1998. “Acquiring Contextual Information for Term Disambiguation”. *Proceedings of Computerm’98*, 86-90.

[15] Nakagawa, H. and Mori, T. (1998). Nested Collocation and Compound Noun For Term Extraction, in *Proc.of Computerm'98*, pp.64-70.

[16] Firth, J. A synopsis of linguistic theory 1930-1955. (1957). *Studies in Linguistic Analysis*, Philological Society, Oxford.

[17] 影浦峡 (2000) 計量情報学. 丸善