タームのrepresentativenessを測るための新指標

2P－3

久光徹† 丹羽芳樹† 辻井潤一‡
†日立製作所中央研究所 ‡東京大学理学部情報科学科

1.はじめに

文書検索において，検索の結果得られた文書数が大きい場合，その内容を把握し，意図した方向へと検索を進めることは容易ではない．これを補助するために，文書集合の内容を俯瞰する特徴語をナヴィゲーションウィンドウに提示する方法が提案されてきた(Niwa 1997)．本研究は，与えられた文書集合中から特徴語を選ぶために，単語の話題性もしくは分野代表性(representativeness)をはかる新しい指標を提案する．日経新聞を用いた選別能力の評価実験を通して，新指標の有効性を示す．

2. 従来の研究

2.1 従来用いられてきた諸指標

情報検索やターム抽出の分野でも，語の「話題性」や「分野代表性」(すなわちrepresentativeness)を測るための指標が数多く提案されてきた(Kageura et al. 1998)．語の重要度に関する指標は，歴史的には情報検索の分野で語の重み付けのために導入され，最も著名な例は *tf-idf* (Salton et al. 1973)であろう．*idf* は，全文書数N_{total}をある単語wが現れる文書数$N(w)$で割ったものの対数，*tf*は単語wの文書d内での出現頻度$f(w, d)$であり，*tf-idf*は，これらの積として，$f(w,d)\times\log(N_{total}/N(W))$で与えられる．さまざまな変形があるが，*tf-idf* の基本的な性質として，「単語がより多く，より少ない文書に偏って出現するほど大きくなる」ように設定される．上記文献には記述されていないが，この指標を特定の文書中での単語の重要度でなく，文書集合全体での単語の重要度を測る指標に拡張する自然な方法は，$f(w,d)$を，wの全文書中での頻度$f(w)$に置き換えることである．4.1ではこの平方を取った$f(w)^{1/2}\times\log(N_{total}/N(W))$を用いる．

他にも，注目する単語の，与えられた文書カテゴリごとの出現頻度の差異の偶然性を測り，偶然でない度合いが高いものをχ^2検定で測る手法(長尾 他, 1976)の他，隣り合う単語の共起の強さを様々な指標で測る手法が提案されている(Cohen 1995, Kita et al. 1994, Franzi et al. 1996, Nakagawa et al., 1997).

2.2 問題点

上記の各指標には，我々の目的に応用するには以下の問題があった：

(1) *tf-idf* (及びその類似手法)の精度は，経験上語の頻度の寄与が大きすぎ，「する」のような高頻度不用語の排除ができにくい．
(2)特定の語のカテゴリ間での分布の違いを比較する方法は，用途が限定される．
(3)隣り合う単語の共起の強さを利用する手法では，1単語ごとの重要度が評価できない．
(4)従来は重要／非重要を分ける閾値の設定が困難かつ恣意的になりがちであった．

本報の目的はこのような問題の無い指標を与える事である．

3. "representativeness"を測るための新指標

3.1 基本方針

あるターム(単語または単語列)が"representative"であるとは，そのタームがある話題(もしくは複数の話題群)を想起させてくれることを指す．この性質は，検索の結果得られた文書集合の内容を俯瞰し，新たにキーとなるタームを示唆する際に重要である．

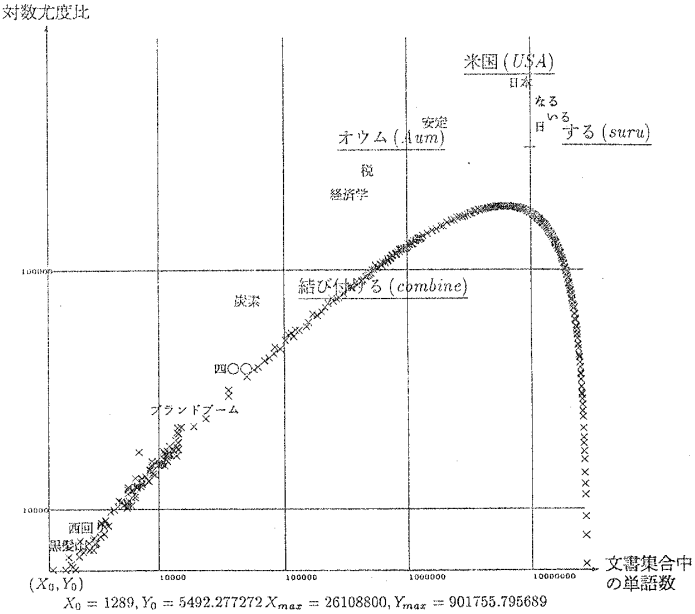
このような性質を測る際の基本的な考え方として，"*You shall know a word by the company it keeps*".(Firth 1957)がある．本報では，これを数学的に言い替えることにより，タームのrepresentativenessを測る指標を導入する．すなわち，

W: ターム
$D(W)$: Wを含む文書すべての集合
D_0: 全文書の集合
$P_{D(W)}$: $D(W)$における単語分布
P_0: D_0 における単語分布

とするとき，Wのrepresentativeness $Rep(W)$を，2つの分布$\{P_{D(W)}, P_0\}$の距離 $Dist\{P_{D(W)}, P_0\}$に基づいて定義する．

単語分布間の距離の計測方法としては，比較実験の結果対数尤度比(log-likelihood ratio)を用いた．

図1は，日経新聞1996年版の記事を用い，そこにあらわれるいくつかの語Wに対し，各語Wについて，$D(W)$の含む単語数 $\#D(W)$を横軸に，$Dist\{P_{D(W)}, P_0\}$を縦軸にプロットしたものである．図から見られるとおり，$\#D(W)$が近いターム同士で比較すれば，たとえば「米国」は「する」，「オウム」は「結び付ける」より$Dist\{P_{D(W)}, P_0\}$の値が高く直感と合致する．しかし，$Dist\{P_{D(W)}, P_0\}$は，$\#D(W)$が大きくなるにつれて増加するため，このままでは$\#D(W)$が離れたターム同士のrepresentativenessを適切に比較することはできない．実際，「オウム」は「する」と$Dist\{P_{D(W)}, P_0\}$の値が同程度となり，直感に合致しない。



$X_0 = 1289, Y_0 = 5492.277272\, X_{max} = 26108800, Y_{max} = 901755.795689$

図1 $Dist(P_{D(W)}, P_0)$ と $Dist(P_D, P_0)$

A New Measure for Representativeness of a Term
Toru Hisamitsu,† Yoshiki Niwa,† and Jun-ichi Tsujii‡
† Central Research Laboratory, Hitachi, Ltd.
‡ Department of Information Science, The University of Tokyo

3.2 距離の正規化

そこで，さまざまな数の文書をランダムサンプリングし，その結果得られた文書集合Dに対して$(\#D, Dist\{P_D, P_0\})$を計算し，図2に"×"を用いてプロットした。これらの点は，(0, 0)に始まり$(\#D_0, 0)$に終わる一つのなめらかな曲線により良く近似できる．以下，この曲線をベースライン曲線と呼ぶ．ベースライン曲線を指数関数を用いた近似関数$B(\cdot)$で近似し，距離を$B(\cdot)$で正規化した値：

$$Rep(W) = Dist\{P_{D(W)}, P_0\}/B(\#D(W))$$

によりWのrepresentativenessを定義する．

ランダムサンプリングした文書集合Dにおける$Dist\{P_D, P_0\}/B(\#D)$は，さまざまなコーパスにおいて，安定して平均Avrがほぼ1，標準偏差σが0.05程度であり，最大値が Avr + 4σ を越えることはなかったので，$Rep(\cdot)$の値が「意味のある値である」と判断するための閾値として，Avr + 4σ =1.20を設けた．

4. 実験結果

4.1 新聞記事中のモノグラムに関する実験

日経新聞1996年版の記事中，総頻度が3以上の単語から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」「好ましくない d」の3種類に人手で分類した．上記20,000語を何らかの方法でソートしたときに，各クラスに分類された語の，先頭からN位までの累積出現頻度グラフを比較する．比較の対象として，ランダムソート，頻度，2.1で述べた全文書を対象とした*tf-idf*の変形版を用いた．

4.2 結果

図2は，分類が"a"となったものの累積頻度を，ランダム，頻度，*tf-idf*，新指標のそれぞれを用いた場合で比較したものである．グラフから明らかに，ランダム＜頻度＜*tf-idf*＜新指標の順で「好ましい」と分類される語の優先順位を上げる力が強い．改善は(劇的ではないが)有為である．*tf-idf*は，重要語抽出においては上回ることがかなり困難な指標であるため(Caraballo et al. 1999)，充分肯定的な結果といえる．

図3は，分類が"d"となったものの累積頻度の比較であり，新指標の選別能力の優位性がより際立っている．頻度と*tf-idf*はランダムな場合と変わらず，「不要語」特定能力の低さを現している．

以上の結果を細かく調べると，競馬馬の名前や力士の名前などの特定のジャンルの固有名詞のRep値が高い．これは，比較的均質な一般の記事集合と，「相撲の勝敗」，「競馬の勝敗」などのような，形式，出現単語，文体等が全く異質な記事が混在しており，そのような部分集合に属する単語にバイアスがかかるためと思われる．このようなものを除いた場合にどうなるかを調べるのは興味深い．

5. まとめ

本報では，タームの"representativeness"をはかるための指標を導入した．この指標は，(1) 数学的な意味付けが明瞭であり，(2)高頻度タームと低頻度タームの比較が自然にでき，(3) 閾値の設定が自然にでき，(4) 任意の長さのタームに対して適用できる，等の性質を持つ．実験によれば，提案指標は不要語の同定にとりわけ有効であり，「高頻度かつrepresentativenessの低い語」を選ぶことによる，stop-word listの自動作成や，文献類似度計算における語の重み付けの精度改善等への応用が期待される．

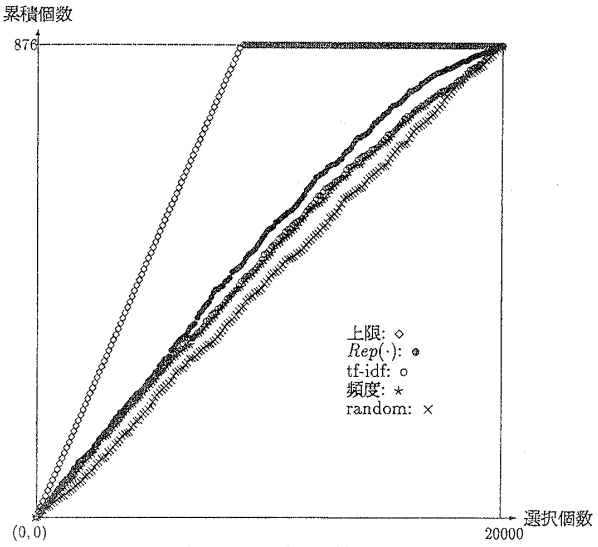


図2 Class-a に属する単語のソーティング

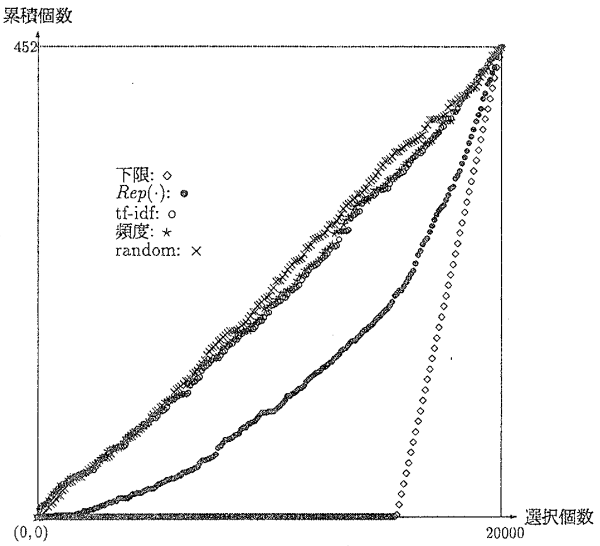


図3 Class-d に属する単語のソーティング

参考文献

Caraballo, S. A., Charniak, E. (1999). Determining the specificity of nouns from text. *Proc. of WVLC'99* .

Cohen, J. D. (1995). Highlights: Language- and Domain-independent Automatic Indexing Terms for Abstracting, J. of American Soc. for Information Science 46(3), pp.162-174.

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

Frantzi, K. T., and Ananiadou, S., and Tsujii, J. (1996). Extracting Terminological Expressions, *IPSJ Technical Report of SIGNL*, NL112-12, pp.83-88.

Kageura, K. and Umeno, B. (1998). Methods of automatic term recognition: A review. *Terminology* 3(2), pp.259-289.

長尾真, 水谷幹男, 池田浩之 (1976). 日本語文献における専門用語の自動抽出, 情処論文誌 17(2), pp.110-117.

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

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

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