組み合わせ確率を用いた特徴単語選択方法

Topic Word Selection Based on Combinatorial Probability

(株)日立製作所 中央研究所
久光徹　丹羽芳樹

1. はじめに

与えられた文書集合を特徴付ける単語を選出する技術は，情報検索や情報抽出における基盤技術の一つである。類似文書検索や文書分類においては，処理の第 1 段階として，検索キーとなる文書(群)や，分類すべき文書(群)に含まれる単語集合から，ノイズ除去による精度向上のために特徴単語を選択することが多く，文書検索インタフェイスにおいても，検索の結果得られた文書集合の内容を概観できる語を表示することは有用である[1]。我々は，[2]において，特徴単語選択のための，組み合わせ確率に基づく単語の重み付け指標を提案した。本報告では，既存の 7 指標との比較を含む，より詳細な評価結果について述べる。

2. 重み付けの方法

以下では，「文書集合 D を特徴付ける単語」を，「D 中に特異的に多く現れる単語」と解釈する立場に立つ。D は，任意の文書集合であってよいが，ここでは，文書検索の結果得られた文書集合の内容概観を与えることを想定し，D として，与えられた単語 w を含む文書全体の集合 $D(w)$ を考える。$D(w)$中の単語数は，異なり数でしばしば数千個を超えるため，指標は効率的に計算できるものでなければならない。

2.1 従来の重み付け指標

上述の効率性への要求を考慮すると，候補となる指標は比較的単純なものに限定される。本節では，提案する指標と比較の対象とした 7 種類の重み付け指標を示す。以下，v は $D(w)$ 中の任意の単語を表す。

2.1.1 ヒューリスティックな指標

・***tf***
詳しくは $tf(v|D(w))$。v の $D(w)$での頻度。

・***tf-idf***
Salton らによって提案された指標[3]で，
$tf\text{-}idf(v|D(w))=tf(v|D(w))\times\log(N_{all}/N(v))$
で定義する。ここで，N_{all}は全文書数，$N(v)$は v が現れる文書数。

・***tf/TF***
$tf/TF = tf(v|D(w))/TF(v)$，但し，$TF(v)$は v の全文書集合中での頻度。tf/TFは，v の $D(w)$ 中での出現確率と，全文書中での出現確率とを比較したものである。

・**SMART**
情報検索の分野で近年提案されたもので[4]，この重みに対して最適化された文書類似度計算方法とともに用いると，高精度な類似文書検索ができるとされている。

$$SMART(v)=\left\{\sum_{d\in D(w)}\frac{\log(tf(v|d))+1}{Ave_{u\in d}\{\log(tf(u|d))+1\}}\right\}\times\log\frac{N_{all}}{N(v)},$$

ここで，$Ave\{\bullet\}$は，{}内の要素の平均を取るオペレータ。SMART は，*tf-idf* を文書長に関して正規化し，精緻化したものである。

2.1.2 確率的な指標

2.1.1 で示したヒューリスティックに定義された指標に対し，数学的に健全な意味を持つ指標が存在する。ここでは，提案指標と比較するため，3 種の代表的な確率的指標について検討した。すなわち，対数尤度比(Log-likelihood Ratio; 以下，**LLR** と呼ぶ)，χ^2値(Chi-squared value; 以下 **CS** と呼ぶ)，Yates 補正した χ^2値(以下 **CS2** と呼ぶ)である。これらはすべて，「$D(w)$内外で単語 v の分布が異なる」という仮説の自然さを測るものある。

対数尤度比を用いる場合，「$D(w)$内外で単語 v の分布が独立」という仮説と，「$D(w)$内外で単語 v の分布は等しい」という二つの仮説に基づいてパラメータを最尤推定する。それぞれの仮説に基づいて得られたパラメータを用いて観測事象の生起確率を計算し，その比の対数をとったものが対数尤度比である。χ^2値は，「$D(w)$内外で単語 v の分布は等しい」という仮説が正しいとした場合の，観測事象の起こりにくさを測るのに利用する。Yates 補正は，低頻度事象をより正確に扱うため，しばしばχ^2値の補正に用いられる方法である。表 1 に，**LLR, CS, CS2** の計算に用いられる 2×2 分割表と，各指標の定義式を示す。この場合の自由度は(2-1)×(2-1)=1 である。

表 1
LLR, CS, CS2 を計算するための 2×2 分割表

	単語 v の延べ数	v 以外の単語の延べ数	計
$D(w)$内	k	$n-k$	n
$D(w)$外	$K-k$	$N-n-K+k$	$N-n$
計	K	$N-K$	N

・**LLR** の定義式

$$k\log\frac{kN}{nK}+(n-k)\log\frac{(n-k)N}{n(N-K)}+(K-k)\log\frac{N(K-k)}{K(N-n)}+(N-K-n+k)\log\frac{N(N-K-n+k)}{(N-K)(N-n)}.$$

・**CS** の定義式

$$\frac{N\{k(N-K-n+k)-(n-k)(K-k)\}^2}{nK(N-K)(N-n)}.$$

・**CS2** の定義式

$$\frac{N\{k(N-K-n+k)-(n-k)(K-k)\}^2}{nK(N-K)(N-n)}$$

$$;if\ k>5\ \&\ n-k>5\ \&\ K-k>5\ \&\ N-K-n+k>5$$

$$\frac{N\{|k(N-K-n+k)-(n-k)(K-k)|-\frac{N}{2}\}^2}{nK(N-K)(N-n)}$$

$$;otherwise$$

2.2 提案する指標(HGS)

[2]で提案した重み付け指標は，次のようなものである。すなわち，全文書の単語数を N, $D(w)$の単語数を n, 単語 v の全文書中での頻度を K, v の $D(w)$中での頻度を k としたとき，v の $D(w)$中での重み $W(N, K, n, k)$を，「N 個の玉の中に K 個の赤い玉があるとき，任意に取り出した n 個の玉の中に赤い玉が k 個以上含まれる確率」(これを $hgs(N, K, n, k)$と書く)の対数値の符号を反転させた値として定義する。ここで，「N 個の玉の中に K 個の赤い玉があるとき，任意に取り出した n 個の玉の中に赤い玉がちょうど k 個含まれる確率」(これを $hg(N, K, n, k)$と書くことにする)は，k を確率変数としたとき超幾何分布(hypergeometric distribution)と呼ばれている。以上を式で書くと次のようになる：

$$W(N,K,n,k)=-\log(hgs(N,K,n,k)),$$

$$hgs(N,K,n,k)=\sum_{l\geq k} hg(N,K,n,l),$$

$$hg(N,k,n,l)=\frac{C(K,l)C(N-K,n-l)}{C(N,n)}$$

$$=\frac{n!K!(N-K)!(N-n)!}{N!l!(n-l)!(K-l)!(N-K-n+l)!}.$$

$$(\min\{0,N+K-n\}\leq l\leq \max\{n,K\})$$

式中，$C(t,u)$は，t 個の異なるものの中から u 個を選ぶ組み合わせの数である。

ここで，「k 個以上」である場合の和をとることの確率的な意味について説明する。指標を，単に $hg(N, K, n, k)$の符号を反転させたものとすると，$k_1<k_2$であって，$hg(N, K, n, k_1)= hg(N, K, n, k_2)$となるような k_1, k_2 が存在しうる。したがって，指標値だけから「出現が特異的に少ない v_1」と，「出現が特異的に多い v_2」を区別することができない。そこで，$k\leq l$ なる l について $hg(N, K, n, l)$の和を取った $hgs(N, K, n, k)$を用いると，この値は，「観測された事象が，"単語 v が出現可能な最大個数(=$min\{n,K\}$)現れる"という事象からどの程度乖離しているか」を表し，「v の出現が特異的に多いこと」を測るための適切な指標となる。

提案する重み付け指標を，以下では，便宜上 **HGS** と呼ぶ事にする。**HGS** の計算では，各単語の出現に独立性を仮定するため，この単純化による精度の限界が存在するはずである。しかし，同じ仮定に立つ **CS** や **CS2** と異なり，パラメトリックモデルによる近似を行わないため，独立性を仮定する確率モデル本来の精度を引き出しているはずであり，**CS** や **CS2** より効果的であることが期待できる。

2.3 重みの効率的な計算方法

定義式だけからは，**HGS** の計算が効率的にできることは必ずしも自明ではない。実際に **HGS** を計算する際は，2.2 の定義式における $hg(N, K, n, l)$の計算に際して，まず対数を取り積和変換する。階乗 $t!$の計算は，$t<150$ のとき表を引き，そうでないときは Stirling の公式で近似する。こうすることにより，二項分布によるパラメトリック近似を行わずとも，任意の(N, K, n, l)に対して超幾何分布の確率が高精度に計算可能である。$hgs(N, K, n, k)$を求める際は，各項の比を取ることによって和の収束性を調べ，収束が早い場合は少ない項数で切り上げる等の工夫をする。また，「特異的に出現が多い」ものを求めるのが目的なので，$hg(N, K, n, k+1) > hg(N, K, n, k)$の時は直ちに計算をやめ，$W(N, K, n, k)$として $\log(hg(N, K, n, k))$を返す。これは負値となるが，単語の序列化には利用可能である。これらより，例えば Compaq AlphaServer 8200 (300MHz)上で秒速 10,000 回程度の重み付け計算が可能であるため，既に[1]で述べた検索システムに実装されている。

3．実験

3.1 8 指標の比較

$D(w)$から特徴単語を選択するという目的に関する各指標の効果を調べるため，**HGS** と，2節で挙げた 7 種類の重み付け指標を用いて比較実験を行った。以下，実験手順の説明のため，これらの計 8 種類の指標をまとめて M と書く。日経新聞 1998 年版より，情報検索のキーワードとなりうる，$D(w)$の含む文書数が似通った w を 2 語ずつ，計 8 単語選んだ。8 単語と，各々に対する $D(w)$が含む文書数は次の通り(括弧内の数字が $D(w)$の文書数)：

{エリツィン(947), オリンピック(934), オウム(265),エイズ(202), イントラネット(152),
プリペイドカード(126), オゾン(52),
テポドン(50)}

これらがすべて片仮名語であるのは，形態素解析による誤分割の影響を受けにくいことが理由である。

M の各要素 m により，各 $D(w)$に含まれる全ての単語を重み付けし，それぞれの上位 50 位までとった単語の集合を $w(m, 50)$とし，これらをマージした単語集合を $w(M, 50)$とする。$w(M,$

50)に含まれる各単語に対し，$D(w)$の内容を概観するうえで有用と思われるものに"P"，概観に現れるのにふさわしくないものに"N"，どちらともいえないものに"U"を付与した。以下は，"P" と"N"への分類に際して用いた主な指針である。以下の要件は，必ずしも互いに独立なものではない。

分類"P"：

- $D(w)\cap D(v)$は複数の文書からなり，それらは概ねwに関する特定のトピックを扱う文書の集合となっている。
- $D(w)\cap D(v)$はwに関するトピックを扱う文書(単数または複数)であり，vはその中で中心的な役割を果たしている。
- vとwは，何らかの動詞の格フレーム中に共起している，もしくはそう解釈できる。
  例：「エイズの治療に用いられるAZTは…」
  (→AZTでエイズを治療する)

分類"N"：

- $D(w)\cap D(v)$に含まれる文書中において，vとwは無関係なコンテクストに出現しているに過ぎない。
- vは，「ある」,「いる」の類の不要語，接続詞等の機能語，たまたま同じ並びで複数回出現する数字列など，検索キーワードとして用いることが不自然な単語である。

このようにして$w(M, 50)$中の単語を分類し，各 $m\in M$ に対し，$w(m, 50)$中に $w(M, 50)$で"P"，"N"と分類される単語がそれぞれ何個含まれるかを数え，これをそれぞれの単語と

表2

単語wと指標mについて，$w(m, 50)$中でPと判定された語数とNと判定された語数

	tf		*tf-idf*		*tf/TF*		*SMART*		*LLR*		χ^2		補正$\chi 2$		HGS	
	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N
エリツィン	25	23	38	10	11	36	35	13	26	22	38	9	38	9	39	9
オリンピック	14	36	29	21	4	44	30	20	17	30	47	3	47	3	44	6
オウム	22	26	43	7	11	38	43	6	24	25	49	1	49	1	48	2
エイズ	20	27	38	12	12	33	34	15	21	27	41	7	44	4	46	3
イントラネット	18	27	34	10	17	31	33	11	19	26	24	16	25	15	39	4
プリペイドカード	17	32	34	16	24	19	26	24	17	32	29	17	32	13	43	7
オゾン	12	33	30	13	16	31	29	17	13	32	23	24	32	15	40	4
テポドン	25	21	41	5	27	21	39	8	26	19	36	13	43	6	42	5

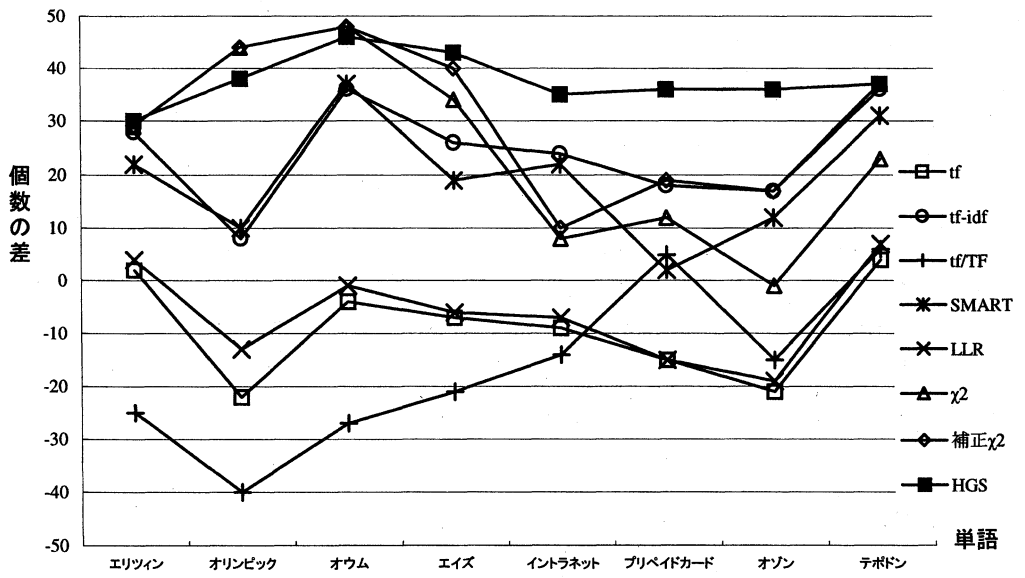

図1

$w(m,50)$中で，Pと判定された単語とNと判定された単語の個数の差

指標の組について示したのが表 2 である。比較を単一の尺度で行うため，それぞれの指標と単語の組みについて，$w(m, 50)$で"P"と分類された単語から"N"と分類された単語数を引いた結果を図示したものが図 1 である。

ヒューリスティックな重み付け指標に対しては，上記 8 単語すべてについて，**HGS** の優位性が示された。確率的な指標に関しては，8 単語中 3 単語で **CS2** が，2 単語で **CS** がそれぞれ僅差で **HGS** を上回ったが，他の単語では **HGS** が上回った。特に半数の単語では **HGS** が大きく上回った。

なお，参考に，w="エイズ"の場合の，各重み付けによる先頭 50 位と，各重みが D("エイズ")の単語に導入する重みの順序相関を調べたところ，*tf-idf* と SMART が非常に類似していること (これは SMART の定義から自然である)，これらの方法では高頻度不要語の排除に難点があることが分かった[2]。

3.2 CS2 と HGS の詳細比較

HGS, CS, CS2 の 3 指標は，ヒューリスティックな指標と異なり，ある単語の出現の「特異的な多さ」に関する"閾値"を，数学的な根拠に基づいてアプリオリに設定できる。例えば，**HGS** の場合，全コーパス中に 1 回だけ現れる単語 v が $D(w)$中に現れる場合の重みであり，**CS, CS2** の場合，危険率 0.1%で「$D(w)$内外で単語 v の分布が等しい」という仮説が棄却できるようなχ^2 値を閾値に設定できる。このような性質においても，性能の面からも，**CS, CS2** と **HGS** は類似性が高いため，**HGS** と，従来指標の中で最も性能が良い **CS2** とをより詳細に比較した結果，以下がわかった。

(1) 8 種類の w について，$D(w)$中の **HGS** で閾値以上となる単語数と，$D(w)$ 中の **CS2** で閾値以上となる単語数を比較したところ，どのケースでも，前者は後者の約 2 分の 1 であった。

(2) 8 種類の w について，$D(w)$中の，**HGS** で閾値以上をとる単語集合と，**CS2** で閾値以上をとる単語集合を比較したところ，前者は後者に包含された。

(3) **CS2** は，**HGS** と比較して，より低頻度の語を上位にランクする傾向がある(例えば，w="エイズ"の場合の上位 50 位の単語の平均 *df* は，HGS が 26.9，CS2 が 19.5)。

(4) w="エイズ"の場合について，$D(w)$中で，**HGS** で閾値以上となる単語集合 A(1092 単語)と，**CS2** で閾値以上となる単語集合 B(2127 単語)を，それぞれランダムにソートした後，100 個ずつランダムサンプリングして，"P"と分類された単語数と，"N"と分類された単語数を調べたところ，A ではそれぞれ 43 個，45 語，B ではそれぞれ 21 語，66 語であった。

(5) w="エイズ"の場合，**CS2** で閾値以上となる単語で，**HGS** で閾値未満となる単語から 100 個ずつ 2 回ランダムサンプリングし，その中で"P"と分類された単語数を調べた結果，それぞれ 6 個，5 個であった。

以上から，次の予想が成り立つ：

HGS と **CS2** が，$D(w)$内でそれぞれの閾値以上となる単語中に"P"と分類される単語を捉える力は，recall ではほぼ等しく，precision では，**HGS** が **CS2** を大きく上回る。

このことを厳密に検証するためには，より多くの単語での実験が必要であるが，概ね，**HGS** は **CS2** に比べ効果的な指標であると判断できる。

4．まとめ

本報告では，文書集合内を特徴付ける単語を選出するための重み付け指標として，超幾何分布を応用した確率値に基づく指標を提案した。すなわち，全文書 D_0 中の単語数を N，w の D_0 中での頻度を K，D の単語数を n，w の D 中での頻度を k としたとき，「N 個の玉の中に K 個の赤い玉があるとき，任意に取り出した n 個の玉の中に赤い玉が k 個以上含まれる確率」の対数値の符号を反転した値である。

D として，キーワード w を含む文書集合 $D(w)$を取り，提案する重み付け指標による先頭 50 位中に，$D(w)$の内容を概観するにふさわしい (及び，ふさわしくない)単語がどの程度出現するか調べたところ，*tf-idf* や SMART を含むヒューリスティックな指標群，対数尤度比や(補正)χ^2値等の従来の確率的指標群に比較して，提案する重み付け指標のほうが，好ましい単語を優先する効果が高いことが分かった。

謝辞

本研究の一部は，IPA 独創的情報技術育成事業の支援を受けて行われました。本研究を進めるにあたり有益なコメントを頂いた，東京大学理学系研究科 辻井潤一教授と，国立情報学研究所 影浦峡助教授に感謝致します。

参考文献

[1] 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*, pp.19-20.

[2] 久光徹 丹羽芳樹 (2000) 組み合わせ的確率モデルに基づく特徴単語選択方法---超幾何分布の応用--- 情処 NL 研報告, Vol. 00-NL-140, pp.85-90.

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

[4] Singhal, A., Buckley, C., and Cochrane, P. A. (1996) Pivoted Document Length Normalization, *Proc. of ACM SIGIR '96*, pp. 126-133.