自然言語処理 140−12
（2000. 11. 21）

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---

久光 徹　丹羽 芳樹
（株）日立製作所 中央研究所

与えられた文書集合を特徴付ける単語を選出することは，様々に応用できる有用技術である。「文書集合を特徴付ける」を，「文書集合中に特異的に多く現れる」と解釈し，これを捉えるために，文書集合 D 中の単語 w に対し，以下の確率値に基づく重み付けを提案する。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」が小さいほど，w に大きな重みを与えるのである。この指標の有効性を，5指標に関する比較実験により示し，併せて上記の確率の効率的計算方法を述べる．

Topic Word Selection Using a Method of Word Weighting Based on Combinatorial Probability
−Use of Hypergeometric Distribution−

Toru Hisamitsu　Yoshiki Niwa
Central Research Laborarory, Hitachi, Ltd.

This paper proposes a method of selecting “characteristic words” from a document set. The selection is done by using the weight that is assigned to each word in the document set. The weight is calculated by using the hypergeometric distribution. A comparative evaluation of five methods of word weighting (including *tf-idf* and SMART) revealed that the proposed method is superior to existing methods. An effiecient method of calculating the hypergeometric probability is also shown.

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ことは，様々に応用できる有用な技術である。例えば，文献検索の結果，キーワード w を含む文書集合 $D(w)$ が得られたとき，その内容の概観を，「$D(w)$ に含まれる特徴的な単語」の集合として提示することは，情報検索インターフェイスにおいて有用であることがわかっているが [1]，そのためには，そのような単語を，適切かつ高速に選択するための重み付け方法が必要である。

「$D(w)$ を特徴付ける単語」を，「$D(w)$ 中に特異的に多く現れる単語」と解釈した場合，単語頻度に基づいてこれを捉えようとする指標は種々考えうるが，$D(w)$ に含まれる文書数はしばしば数百〜千となり，数千個〜1万個を超え

る異なり単語を含む。従って，リアルタイムで$D(w)$中の単語全ての重み付けを行うためには，重み付けが効率的に行えなければならない。

本報告では，上に述べた応用のため，$D(w)$から「特徴語」を抽出するための重み付け方法として，組み合わせ的確率値に基づく重み付け方法を提案する。提案する重みは，超幾何分布を応用した確率計算に基づくものであり，その定義から，高頻度語や低頻度語に偏らない，公正な重み付けが可能であると期待される。

以下，2節で従来の重み付け方法を紹介した後，提案する方法を示す。3節では，2節で示す5種類の重み付け方法の比較実験を行い，提案方法の優位性を示す。4節では，重みの効率的な計算方法について簡単に述べ，5節では，まとめと他のタスクへの応用の可能性を示す。

2. 重み付けの方法

2.1 従来の方法

単語の重み付けに関しては，文献検索のための索引語抽出を目的とする種々の方法が提案されている(レヴューとして，例えば[2]を参照)。1節で述べたリアルタイム性の要求を考慮すると，候補となる方法は，比較的単純なもの(単語の共起情報などは用いない)に限定される。以下では，$D(w)$中の単語vに対する4種類の重み付け手法を示す。

- ***tf***

  詳しくは$tf(v|D(w))$。もっとも単純な重みで，vの$D(w)$での頻度そのものを用いる。

- ***tf-idf***

  Saltonらによって提案された方法で[3]，

  $$tf\text{-}idf(v|D(w))=tf(v|D(w))\times\log(N_{all}/N(v))$$

  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。すなわち，*tf-idf*は，より少ない文書に偏って出現する単語が数多く現れる時，大きくなる．

- ***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。

- **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，最も高精度な類似文書検索ができるとされている。

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

  ここで，$Ave\{\bullet\}$は，{}内の要素の平均を取るオペレータ。

*tf*は，その単純さに関わらず，様々なタスクにおいて，複雑な指標より有効なことが多いことが示されている上，計算コストは最小である。

*tf-idf*も，計算コストが低く，有用な尺度である。しかし，単語頻度に引きずられる傾向が強く，その結果，「する」「いる」などの「不要語」の重みが高くなるため，我々の応用目的においては問題があることがわかっている。

*tf/TF*は，意味が明確であるが，$TF(v)$=1であるようなvが$D(w)$に現れた場合，それらの*tf/TF*値は1となり，*tf-idf*とは逆に，低頻度語の過大評価が生じる問題がある。

SMARTは，その定義からわかるように，*tf-idf*に，文書長に関する正規化を加えて精緻化したものである。計算コストはかなり高い。

2.2 提案方法

2.1で述べたように，従来の手法には，高頻度語，または低頻度語を過大評価する傾向が見られ，またその定義式は，さほど数学的な根拠が

あるものではなかった。唯一，*tf/TF* の意味は明快だが，明らかな低頻度語過大評価の問題がある。

そこで，「特異的に多く現れること」を，数学的な根拠が明快で，しかも低コストで計量する重み付け方法として，本報告では，確率値に基づく重み付けを提案する (以下に示す重みの計算が低コストでできることは自明ではないため，計算方法については 4 節で簡単に触れる)。すなわち，全文書の単語数を N，単語 v の全文書中での頻度を K，$D(w)$の単語数を n，v の $D(w)$中での頻度を k としたとき，「v が特異的に多く出現する」ことを，「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 $hgs(N, K, n, k)$と書くことにする) が低いことと対応付ける。そして，v の重み $W(N, K, n, k)$は，上記の確率の対数値の符号を反転させた値とする。

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と呼ばれている。

以上をまとめると以下のようになる：

$$W(N,K,n,k) = -\log(hgs(N,K,n,k)),$$
$$hgs(N,K,n,k) = \sum_{l \geq k} hg(N,K,n,l),$$
$$hg(N,k,n,l) = \frac{C(K,l)C(N-K,n-l)}{C(N,n)} = \frac{n!K!(N-K)!(N-n)!}{N!l!(n-l)!(K-l)!(N-K-n+l)!}.$$
$$(\min\{0, N+K-n\} \leq l \leq \max\{n,K\})$$

ここで，$C(t,u)$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。実際には，例えば，日経新聞 1998 年分で，w="エイズ"，v="厚生省"のとき，(N, K, n, k)= (26563725, 2792, 47639, 165)，$W(N, K, n, k)$= 613.230484 となる。

ここで，「k 個以上」である場合の和をとるのは，「単語 v が k 個現れる」という事象が，「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離れているかを測るためである。このような和を取ることにより，「出現が特異的に少ない」場合と，「出現が特異的に多い」場合を区別できる。すなわち，$k_1 < k_2$ であって，$hg(N, K, n, k_1) = hg(N, K, n, k_2)$ となる場合，この確率値だけからは，v が k_1 回出現した場合と，k_2 回出現した場合の「珍しさ」の意味の違いがわからない。(図 1)

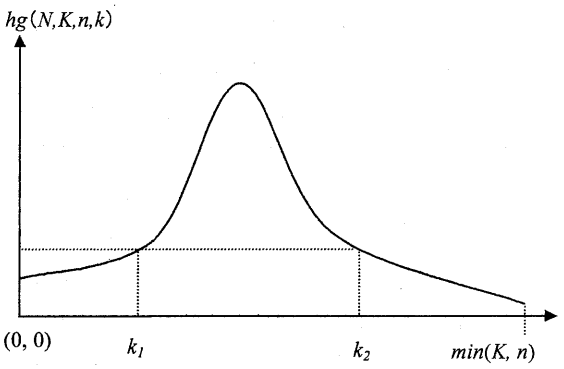

図 1
超幾何分布の確率値自体では区別がつかない場合を示す模式図

しかし，$l \geq k$ なる l について $hg(N, K, n, l)$の和をとることにより，$hgs(N, K, n, k_1) > hgs(N, K, n, k_2)$，したがって $W(N, K, n, k_1) < W(N, K, n, k_2)$ となる。

提案方法は，その確率的解釈により，Nや K と比較して n や k が大きな場合も小さな場合も，一貫した意味付けによる公平な重み付けができる。これは，SMART で行うような文書サイズによる正規化が，確率を使うことにより自動的に行われるからである。提案する重み付け方法を，以下では便宜上 HGS と呼ぶ事にする。

3. 実験

我々の目的に関する HGS の性能を調べるため，2 節で挙げた 5 種類の方法の重み付け精度を比

較した。以下，これら5種類の方法をまとめてMと書く。日経新聞1998年版より，$D(w)$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)}

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。$w(M, 50)$の各要素に対し，各単語があらわれるコンテクストを参照し，$D(w)$の内容を概観するうえで有用と思われるもの(検索内容の確認に有効or内容の絞込みに有効or関連トピックへの手がかりとして有効)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，各$w(m, 50)$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た。

その結果を示したのが図2-(a), (b)である。上記8単語すべてについて，HGSの優位性が示された。このタスクに限っては，SMARTは*tf-idf*よりわずかだが劣った結果となった。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(表1)，各重みがD("エイズ")の単語に導入する重みの順序相関の一部を示した(表2)。これらから，*tf-idf*とSMARTが非常に類似していることと(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窺える。

4. 計算方法についての注意

2.2の定義式における$hg(N, K, n, l)$の計算に際しては，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，二項分布近似を行うことなく，任意の(N, K, n, l)に対して高精度に直接計算可能である。$hgs(N, K, n, k)$を求める際は，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，$hg(N, K, n, k+1) > hg(N, K, n, k)$のときは，直ちに計算をやめて，$W(N, K, n, k)$として$\log(hg(N, K, n, k))$を返す(これは負値)等の工夫をする。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り，既に検索システムに実装されている。

5. まとめと展望

本報告では，与えられた文書集合を特徴付ける単語を選出するための重み付け方法として，文書集合D中の単語wに対し，確率値に基づく重み付けを提案した。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重みを与える。この値は，充分実用的な速度で計算可能(SMARTより高速に計算できる)であるばかりか，*tf-idf*やSMART等による重み付けより有効であった。

上記の重み付け方法は，他にも有用な適用先を持つ。例えば，情報検索の精度評価において，recallとprecisionの異なるシステムの性能を比較する場合，「全部でN文書の中にK文書の正解があるとき，システムがn文書を出力し，その中に正解がk文書あった」場合に$W(N, K, n, k)$を与えて比較すれば，F-measureよりきめ細かい比較ができる。これらについては，稿を改めて報告する予定である。

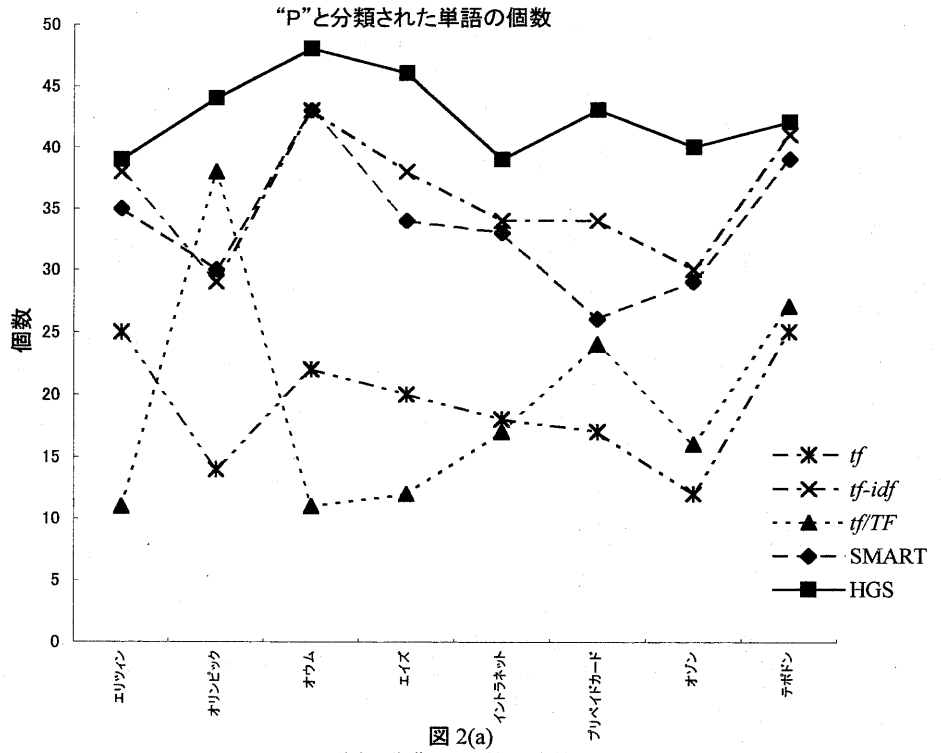

図 2(a)
分類 "P" の語数の比較

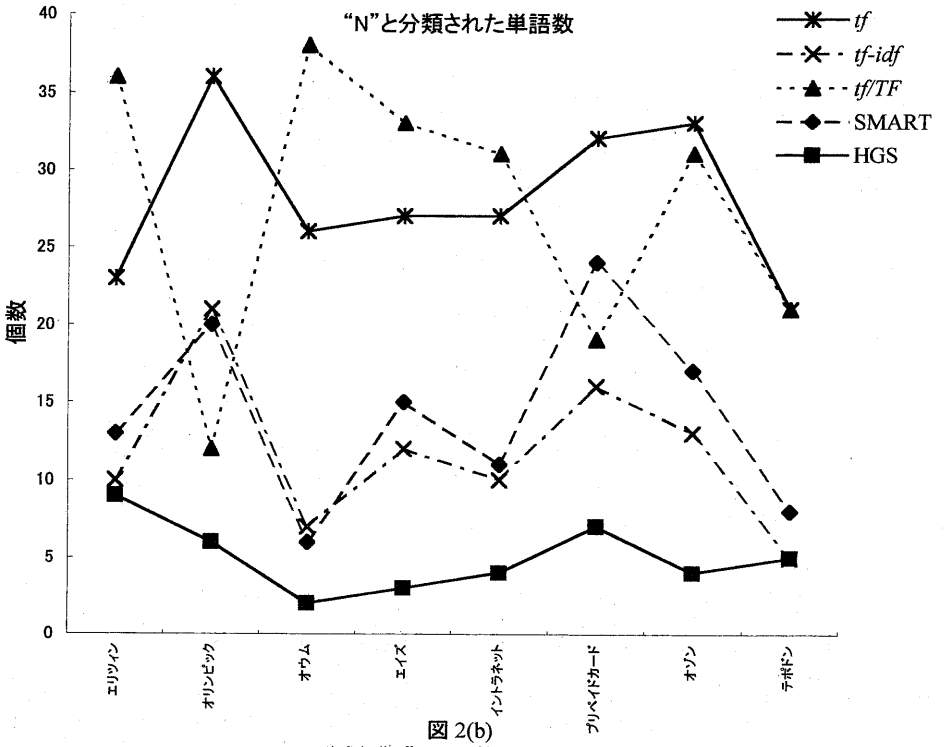

図 2(b)
分類 "N" の語数の比較

表 1
各指標による上位 50 単語（*D*(“エイズ”)より出力）

tf	*tf-idf*	*tf/TF*	SMART	HGS
する	エイズ	六字	エイズ	エイズ
エイズ	薬害	免疫療法剤	薬害	薬害
こと	感染	法廷等の秩序維持に関する法律	感染	感染
ある	薬	平方根	厚生省	薬
年	厚生省	不服申立て	薬	厚生省
人	する	百七十一万	患者	ウイルス
日	患者	筆記体	製剤	製剤
いう	ウイルス	脳せき髄液	血液	血液
なる	製剤	二万二百十五	エイズウイルス	我々
感染	我々	難症	予防	患者
薬害	血液	困難さ	感染症	予防
薬	予防	定三	ウイルス	感染症
月	遺伝子	鶴子	治療	エイズウイルス
一	医療	打切る	医	遺伝子
厚生省	感染症	打ち続く	研究	ワクチン
者	治療	多項式	医療	ミドリ
患者	人	整腸薬	人	治療
この	こと	震わす	製薬	吉富
できる	ワクチン	松石	問題	輸血
ない	研究	主尋問	遺伝子	医療
我々	医	七千八百十九	いう	医
問題	検査	自慰	薬品	インフルエンザ
情報	輸血	氏姓	ある	副作用
日本	エイズウイルス	四則	こと	班
研究	ある	四千二百八十八	事件	及ぶ
よる	ミドリ	三万一千百三十	ミドリ	製薬
医療	薬品	三平方の定理	年	根絶
これ	吉富	三千七百六	輸血	避妊
二	情報	三千七百四十三	検査	投与
同	行政	三千三百四十万	者	薬品
ウイルス	製薬	三千五百四十六	臨床	検査
もの	及ぶ	三千五十四	この	医師
血液	いう	左寄り	ない	臨床
中	問題	困る	する	テープ
氏	医師	合薬	医師	抗体
三	年	公吏	ワクチン	献血
ら	者	後天性	班	発病
予防	副作用	五万二千九百六十	副作用	免疫
治療	公開	貢献し	できる	血友病
行政	インフルエンザ	軽挙	抗体	行政
その	臨床	逆転写	投与	学校医
開発	班	詰めこみ	行政	新薬
製剤	法	疑いぶかい	訴訟	研究
機関	承認	監置	性	承認
遺伝子	投与	感化院	テープ	十字
法	事件	家捜し	ら	安部
医	この	化学式	よる	肝炎
的	性	押売り	なる	碑
の	テープ	延展	日本	原告
米	避妊	一万六千六百八十五	その	郡司

表 2
D(“エイズ”)中の単語に導入される順位の間の Spearman の順序相関

	tf	*tf-idf*	SMART	HGS
tf	1	0.847389	0.816273	0.157063
tf-idf	—	1	0.967031	0.503403
SMART	—	—	1	0.482557

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*

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

[3] Salton, G. and Yang, C. S. (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*, pp126-133.