The 30th Symposium on Information Theory
and its Applications (SITA2007)
Kashikojima, Mie, Japan, Nov. 27–30, 2007

テキスト分類問題を対象としたベクトル空間における距離構造の漸近解析に関する一考察

Statistical Asymptotic Analysis of Distance Measure on the Vector Space in Text Classification Problems

後藤 正幸 * 石田 崇 † 鈴木 誠 ‡ 平澤 茂一 †
Masayuki GOTO, Takashi ISHIDA Makoto SUZUKI Shigeichi HIRASAWA

Abstract— In this paper, the document classification problems in text mining are considered from the viewpoint of statistics model. In the problem in text mining, the several heuristics are applied to practical analysis because of its experimental effectiveness in many case studies. The theoretical explanation about the performance in text mining techniques is required and such thinking will give us very clear idea.

Keywords— information retrieval, tf-idf, document classfication

1 はじめに

近年，インターネットの普及により膨大なテキストデータからの知識発見を扱うテキストマイニングの技法が注目されている [1]．本研究では，テキストマイニングが取り扱う問題の中でも，特に文書分類の問題を取り上げ，形態素解析後の単語の出現分布としてある確率モデルのクラスを仮定し，文書分類のために用いられる距離について分析を行う．一般に，単語の出現頻度に基づく文書分類においては，個々の単語の生起頻度は少なく，多くの単語の頻度がゼロとなってしまうというスパースネスの問題がある．すなわち，このベクトル空間上で一つの文書を表す点は，ゼロを多くの要素に持つベクトルで表現される．しかし,「このような状況で，文書同士の距離による分類がある程度の分類性能を示すのは何故か」という疑問については依然として経験的な解釈が与えられているのみである．著者らは，その理論的根拠を与えるため，各要素の出現頻度を有限に保ったまま，次元数を無限大とする新たな漸近論の概念を導入することにより，解析的に示している [2]．本研究では，この枠組みを一般化し，文書クラスを特徴付ける単語群が複数存在し，かつ分類に無意味となる単語群も存在する場合について漸近性能を示すと共に，その結果について考察を与える．

* 224-0015 横浜市都筑区牛久保西 3-3-1, 武蔵工業大学 環境情報学部 (Musashi Institute of Technology, Fac. of Environmental and Information Studies), E-mail: goto@yc.musashi-tech.ac.jp
† 早稲田大学 理工学部（Waseda University, School of Science and Engineering）
‡ 湘南工科大学 工学部（Syonan Institute of Technology, Fac. of Engineering）

2 文書モデル

本節では，形態素解析により，各文書 について単語への切り出しが行われた後，情報検索やテキスト分類の問題を取り扱い易い問題に落とし込んだモデルであるベクトル空間モデルについて述べる．ベクトル空間モデルでは，文書中の出現単語の頻度に基づき，文書の特徴量を 1 つのベクトルで表現することで，文書を空間上の点として表す．出現単語に基づくベクトル空間を構成し，文書を空間上の点として表現することで，文書同士の類似性を距離の概念によって数学的に取り扱うことが可能である．このモデルは，計算機で実装する際に強力な枠組みを与えるものである．

2.1 ベクトル空間と文書－単語行列

分析対象である文書集合を $\Delta = \{d_1, d_2, \cdots, d_D\}$ とする．Δ 内の全ての文書について，文書内に含まれる単語を抽出する．この単語抽出には，通常，文書の分類や検索のために有効となる単語（有効語）を選定して抽出する．すなわち，助詞や句読点など，文書の内容にあまり関係なく出現する語は分類や検索には意味をなさないため除外する．通常は，有効語として名詞や動詞の語幹の中から全文書中での頻度を考慮して選定される．全文書から抽出された有効語の集合を $\Sigma = \{w_1, w_2, \cdots, w_W\}$ とすれば，各文書の特徴ベクトルを各特長語の出現頻度に応じて，W 次元ベクトルで表現することができる．すなわち，文書集合 Δ から得られる全有効語によってベクトル空間が構成され，文書 d_i を次式で表現することができる．

$$d_i = (v_{i1}, v_{i2}, \cdots, v_{iW})^T \tag{1}$$

ただし，T は転置を表す．ここで，この文書ベクトルを集めた行列

$$A = (d_1, d_2, \cdots, d_D)^T \tag{2}$$

を文書－単語行列 (document word matrix) と呼ぶ．

本稿では触れないが，この特徴ベクトルに含まれるノイズを除去し，意味のある空間において分析を行うための方法として Latent Semantic Indexing という方法が提案されており，この方法ではこの文書－単語行列を特

異値分解することによって得られる主成分空間上でベクトル空間を構成する．

2.2 TF・IDF Measuare

前節において，各文書 d_i のベクトル表現を与えたが，各要素の値を如何に決めるかという問題が残っている．最も簡単な方法として，各単語の出現頻度とする方法がある．k_j^i を文書 d_i に含まれる単語 w_j の出現頻度とし，$v_j^i = k_j^i$，すなわち

$$d_i = (k_1^i, k_2^i, \cdots, k_W^i)^T \tag{3}$$

とする方法であるが，しばしば検索や分類性能が，多くの文書に出現する単語に大きく影響されてしまうという問題がある．いま，k_{w_i} を全ての文書中の単語 w_i の頻度，k_{d_j} を文書 d_j 内の全単語の総頻度，F を全文書中の全単語の総頻度とする．すなわち，

$$F = \sum_{w_i} \sum_{d_j} k_j^i = \sum_{d_j} k_{d_j} = \sum_{w_i} k_{w_i} \tag{4}$$

の関係があるとする．v_j^i として相対頻度を考え，$v_j^i = \frac{k_j^i}{F}$ とする方法や，文書の長さによる影響を解消するために $v_j^i = \frac{k_j^i}{k_{d_j}}$ とする方法もある．

通常，全ての文書にまんべんなく表れる単語は，文書の特徴を規定するためにはあまり意味がない．むしろ，少数の文書において集中的に表れる単語は分類や検索に有効である．そこで，各単語の出現頻度だけでなく，全文章中でその単語が現れる割合を考慮した特長量の算出が必要であり，そのための方法が TF・IDF measure である [1]．TF は Term Frequency の略であり，文字通り単語の出現頻度を表す．一方，IDF は Inverse Document Frequency の略であり，全文書中の単語の出現割合の減少関数を表す．ここでは，TF を文書 d_i における単語 w_j の相対頻度とし，$tf(d_i, w_j) = \frac{k_j^i}{F}$ とおく．IDF は単語 w_j を含む文書の数を $df(w_j)$ とすると，

$$idf(w_j) = \log \frac{D}{df(w_j)} \tag{5}$$

のような関数で定義される．このとき，文書 d_i における単語 w_j の特徴量 v_j^i は，

$$v_j^i = tf(d_i, w_j) \cdot idf(w_j) \tag{6}$$

で与えられる．最近では，TF・IDF measure の情報理論的な解釈についても研究が行われている．

Aizawa[3] は，

$$v_j^i = \frac{k_{w_i}}{F} \sum_{d_j} \frac{k_j^i}{k_{w_i}} \log \frac{\frac{k_j^i}{k_{w_i}}}{\frac{k_{d_j}}{F}} \tag{7}$$

を KL-情報量を用いた TF・KLI measure として提案している．

2.3 文書間の類似度判定

各文書の特徴量がベクトル表現されれば，文書 d_i と文書 d_k の類似度（内容的近さ）は，これらの距離を使って測ることができる．この距離には，ユークリッド距離や内積を用いることも可能であるが，これらの距離は原点付近の 2 点が近いものであると判定する．ほとんどの単語の特徴量が 0 に近い文書同士は内容的に類似しているとは言えないが，ユークリッド距離や内積によれば類似していると判定してしまう．そこで，文書ベクトル d_i と文書ベクトル d_k の余弦をとって類似度とする方法が一般的である．

$$sim_c(d_i, d_k) = \frac{d_i^T d_k}{\|d_i\| \, \|d_k\|} \tag{8}$$

文書検索の問題においては，検索語を特徴ベクトル（クエリベクトルと呼ぶ）で表現し，このクエリベクトルと類似度の高い文書を検索結果として提示する．類似度の高いものからリスト表示することにより，検索結果のランキング機能も有していることになる．文書の類似性評価については，様々な問題に対して，問題の特性を考慮した方法が研究されている．また，もし全ての i について d_i が確率分布を表すベクトルになっている，すなわち $v_j^i = \hat{q}_j^i$ かつ $\sum_{j=1}^W \hat{q}_j^i = 1$ であれば，KL 情報量を用いた距離尺度を使うこともできる．

$$sim_d(d_i, d_k) = \sum_{j=1}^{W} \hat{q}_j^i \log \frac{\hat{q}_j^i}{\hat{q}_j^k} \tag{9}$$

通常，確率モデル同士の距離であれば Divergence を用いた方が整合性がありそうであるが，文書分類や文書検索の分野では情報量のような距離はあまり用いられていない．これは経験的に分類性能が悪いとされているためである．

3 確率モデルの定義

従来のパラメトリックな確率モデルでは，パラメータの次元を固定としたもとで，データ数を増加させた時の挙動が論じられている．本稿では，個々の要素のデータ数は有限である状況で，ベクトルの次元数 W を $W \to \infty$ とすることによる漸近論を論じることにより，スパースネスの問題を取り扱う．本節では，このような状況設定のための確率モデルの定義を示す．

通常，多くの単語はそれぞれ意味が類似するものがあり，同じ文書クラスでは同程度の出現確率を持ち、かつ文書クラスによって生起確率の異なる単語グループに分類できることが想定できる．このような単語グループが T 個存在し，かつ文書分類には全く影響を与えない単語グループが含まれる場合を考える．すわなち，W 個の単語が $T+1$ 個のグループに別れ，各グループ内の単語

は同じ生起確率を持つものとする．それぞれのグループの単語数を r_1,r_2,$\cdots$,r_T,r_{T+1} とおく．ただし，

$$r_{T+1} = W - \sum_{i=1}^{T} r_i \tag{10}$$

である．

さらに，文書 d_t と文書 d_u はそれぞれ確率分布 p^t と p^u に従って生起する経験分布 $\hat{q}^t$ と $\hat{q}^u$ によって表されるものとする．このとき，$k = 1, 2, \cdots, T+1$ と

$$j = \sum_{i=1}^{k-1} r_i + 1,\ \sum_{i=1}^{k-1} r_i + 2,\ \cdots,\ \sum_{i=1}^{k} r_i$$

に対して，

$$p_j^t = \frac{s_k^t}{r_k}, \qquad p_j^u = \frac{s_k^u}{r_k} \tag{11}$$

とおく．通常，文書モデルでは，文書の分類に全く意味を与えない不要語が含まれることを想定する必要がある．そこで，第 $T+1$ カテゴリの単語群は，そのような不要語の集合であるものとし，

$$p_{T+1}^t = p_{T+1}^u \tag{12}$$

とする．

これらの確率が既知であれば，

$$sim_d^*(p^t, p^u) = \sum_{j=1}^{W} p_j^t \log \frac{p_j^t}{p_j^u} \tag{13}$$

$$sim_c^*(p^t, p^u) = \frac{\sum_{j=1}^{W} p_j^t p_j^u}{\sqrt{\sum_{j=1}^{W} (p_j^t)^2}\sqrt{\sum_{j=1}^{W} (p_j^u)^2}} \tag{14}$$

は計算可能である．しかし，文書分類の問題においては，推定量同士で距離を測っていることを想定する方が現実的である．

4 経験分布同士の距離・類似度

ここでは，2 つのクラス C_t と C_u を想定した 2 クラス問題を考える．各文書と単語の出現確率は独立であり，確率 p^t をクラス C_t から出現する文書 d_t の出現確率分布，確率 p^u をクラス C_u から出現する文書 d_u の出現確率分布とする．

そこで，両クラスから出現した単語ベクトルの統計量として $\hat{q}^t = \frac{k_j^t}{k_{d_t}}$，$\hat{q}^u = \frac{k_j^u}{k_{d_u}}$ を考える．ここで両者の総出現単語数は同じく $N = k_{d_t} = k_{d_u}$ と仮定する．

$$sim_d(\hat{q}^t, \hat{q}^u) = \sum_{j=1}^{W} \hat{q}_j^t \log \frac{\hat{q}_j^t}{\hat{q}_j^u} \tag{15}$$

$$sim_s(\hat{q}^t, \hat{q}^u) = \frac{\sum_{j=1}^{W} \hat{q}_j^t \hat{q}_j^u}{\sqrt{\sum_{j=1}^{W} \hat{q}_j^t}\sqrt{\sum_{j=1}^{W} \hat{q}_j^u}} \tag{16}$$

もし真の確率分布が分かっていれば (13) 式，(14) 式で測りたいが，真が分からないので上の距離の推定量を使うことになると考える．

通常の文書分類問題では，単語の頻度に比べてベクトル空間の次元がかなり大きく，スパースネスの問題を有している．特定の統計量が漸近的にどのように真のパラメータに収束するかを議論するよりも，むしろ有限サンプルの統計量をたくさん集めた時にどのようなパフォーマンスが得られるかを議論することが，文書分類の問題の振る舞いを解析する事につながる．

そのため，各単語グループの個数 r_1,r_2,$\cdots$,r_T,r_{T+1} のレートが一定であるという仮定をおき，$W \to \infty$ を考え，

$$R_1 = \frac{r_1}{W},\ R_2 = \frac{r_2}{W},\ R_3 = \frac{r_3}{W},\ \cdots,\ R_{T+1} = \frac{r_{T+1}}{W} \tag{17}$$

とおく．これら R_1,R_2,$\cdots$, R_T, R_{T+1} は によらない定数である．さらに，単語の種類が増えると，それに比例して全単語の出現回数は増えるものと考えられる．そこで，$N = nW$ とする（n は定数）．以上の設定のもとで，$W \to \infty$ とするとき，文書間の類似度について，次の定理が得られる．

定理 1 *n, R_1,R_2,$\cdots$, R_T, R_{T+1} を固定したもとで，$W \to \infty$ としたとき，次の概収束が成り立つ．*

$$sim_c(\hat{q}^t, \hat{q}^u) \to sim_c^*(p^t, p^u), \qquad a.s. \tag{18}$$

□

定理 2 *n, R_1,R_2,$\cdots$, R_T, R_{T+1} を固定したもとで，$W \to \infty$ としたとき，次の概収束が成り立つ．*

$$sim_d(\hat{q}^t, \hat{q}^u) \to \sum_{j=1}^{T+1} R_j \mu_j, \qquad a.s. \tag{19}$$

ただし，μ_j は

$$\mu_j = E_{p^t}\left[\frac{k_j^t}{n} \log \frac{k_j^t/n}{k_j^u/n}\right] \tag{20}$$

をみたす値であり，E_{p^t} は確率分布 p^t による期待値を表す．□

これらの定理より，一般によく用いられる余弦を用いた類似度では，個々の単語の出現頻度が低い場合でも，あるカテゴリで比較的頻出する単語の種類を増やしていくことにより，正しい文書ベクトルの距離を推定できることが示唆される．一方，KL 情報量は，このようなスパースネスの問題を持つ問題についてはうまく働かない．これはこれまでの経験的な知見を裏付ける結果と言える．情報理論や統計的検定の枠組みでは本質的である“ 情報量 ”が本質的にうまく動作しないことは当然とみなすこ

とも出来るが，逆にこのような問題設定において本質的な“ 情報量の概念 ”を構築できるか否かは興味深い課題である．

5 IDF Measure に関する議論

ここでは，文書検索や文書分類の応用においてよく利用される IDF measure について議論を行う．式 (6) の TF・IDF measure にもあるように，IDF measure は，分類や検索に対し，その単語の重要性を表す重みを解釈することができる．すなわち，より重要な語には大きな重みを，重要ではない語には小さな重みを付与してから，文書ベクトル間の重み付け距離を測定することにより，検索性能や分類性能が向上するというものである．

いま，確率 p^t (クラス C_t) から出現する文書 d_1,d_2,$\cdots$,d_g，確率 p^u (クラス C_u) から出現する文書 d_{g+1},d_{g+2},$\cdots$,d_{g+h} の経験分布をそれぞれ，$\hat{q}^{t_1}$, $\hat{q}^{t_2}$, $\cdots$, $\hat{q}^{t_g}$, $\hat{q}^{u_1}$, $\hat{q}^{u_2}$, $\cdots$, $\hat{q}^{u_h}$ と記述する．このとき，$D = g + h$ である．

ここで，文書数 D を多くできたときの IDF measure の振る舞いを考えてみる．前節の議論と同様に，

$$T = \frac{g}{D}, \quad 1 - T = \frac{h}{D}$$

を一定に保ちつつ，$D \to \infty$ とする漸近操作を考えることは，クラス C_t と C_u の双方から文書サンプルが得られていく状況を想定しており，現実的と考えられる．このとき，3 章の式 (11)，(12) などで定義した確率モデルに対し，次の結果が導かれる．

定理 3 W を固定とする．T を固定したもとで，$D \to \infty$ としたとき，第 k グループの単語 w_j に対し，次の概収束が成り立つ ($k = 1,2,\cdots,T+1$).

$$\begin{aligned} idf(w_j) &= -\log \frac{df(w_j)}{D} \\ &\to -\log \{TK_t + (1-T)K_u\}, \quad a.s. \end{aligned}$$

ただし，

$$K_t = 1 - \left(1 - \frac{s_k^t}{r_k}\right)^N, \quad K_u = 1 - \left(1 - \frac{s_k^u}{r_k}\right)^N$$

である． □

この定理の意味するところを考察してみよう．ここでは，簡単のため，各クラスから文書サンプルは同数与えられるものとする．すなわち $T = 1/2$ である．このとき，

$$idf(w_j) \to -\log \left[1 - \frac{1}{2}\left\{\left(1 - \frac{s_k^t}{r_k}\right)^N + \left(1 - \frac{s_k^u}{r_k}\right)^N\right\}\right]$$

となる．通常，s_k^t と s_k^u の差が大きいほど，文書の特徴を現すためには適している単語であると考えられるが，IDF measure は，少なくとも本稿で定義した確率モデルの上では，そのような性質を有していない．この measure は，s_k^t と s_k^u が共に小さい場合に大きなウェイトを与える．

一方，$T >> 1 - T$ の状況を考えてみる．このとき，

$$idf(w_j) \ \sim \ -\log \left\{1 - \left(1 - \frac{s_k^t}{r_k}\right)^N\right\}$$

と s_k^t が主要項になる．これは，2 クラスではなく，より多くの確率分布（クラス）から文書が得られるときに成り立つ状況であろう．このとき，$s_k^t \to 0$ であれば，$idf(w_j) \to$ 大となる．逆説的であるが，s_k^u が大きい場合においても，s_k^t が十分小さければ，$idf(w_j) \to$ 大となる．すなわちクラス C_u から出現する文書の単語 w_j の生起確率が大きく，クラス C_u から出現する文書には単語 w_j が現れない場合には，IDF measure によってこのような単語に大きなウェイトを付与できる．

逆に言えば，s_k^t と s_k^u が共に小さい場合についても IDF measure は単語 w_j に大きなウェイトを与えてしまい，このようなウェイト算法は適切に働かないことが予想される．「s_k^t と s_k^u の差が大きいほど，文書の特徴を現す」と考え，このような考え方にそって適切に重要単語に大きなウェイト付けを行うためには，単語がどのクラスから生起したかを反映した方法が必要である．IDF measure は文書の生起クラスが分からない情報検索問題などで適用されるものの，文書分類問題ではより優れた方法が構築できることを意味する．その意味でこの結果は M.Suzuki の提案する蓄積手法 [4] などの有用性を裏付けるものと考えられる．

6 おわりに

本稿では，文書分類の問題について基礎的なモデルを使ってその性質を考察した．文書分類など，テキストマイニングの分野では高次元のモデルを扱うため，相対的にデータ量の少ない問題を扱っており，今後もこのようなモデルの分析は有用な知見を与えてくれるものと考えられる．

参考文献

[1] 金明哲，村上征勝，永田昌明，大津起夫，山西健司: 言語と心理の統計, 岩波書店，(2003)

[2] M.Goto, T.Ishida, S.Hirasawa: Statistical evaluation of measure and distance on document classification problems in text mining ”, *Proceedings of IEEE 7th International Conference on Computer and Information Technology*, (2007)

[3] A.Aizawa: ”An information-theoretic perspective of tf-idf measures”, *Information Processing and Management*, Vol.39, pp.45-65, (2003)

[4] M.Suzuki: ”Text classification based on the bias of word frequency over categories”, *Proceedings of the International Conference on Artificial Intelligence and Applications(AIA)*, pp.400-405, (2006)