The 29th Symposium on Information Theory
and Its Applications (SITA2006)
Hakodate, Hokkaido, Japan, Nov. 28 – Dec. 1, 2006

文書分類モデルの性質に関する一考察

Information Theoretic Consideration of Document Classification

後藤 正幸 *　　平澤 茂一 †
Masayuki GOTO,　　Shigeichi HIRASAWA

Abstract— In this paper, the document classification problems in text mining are considered from the viewpoint of statistics model. By formulation of statistical hypotheses test which is specified in the problem of text mining, some interesting properties can be visualized. 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],[2]．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，伝統的な統計的仮説検定と漸近論の理論的枠組みから得られる性質について分析を行う．とくに，文書のクラス数が 2である場合には，文書分類の問題は仮説検定とほぼ同様の枠組みで捉えることができる．本稿では，文書クラスを特定するために，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，そこから得られる性質を検討することで，文書分類の本質的な挙動について明らかにする．

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．

* 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)

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．f_{ij} を文書 d_i に含まれる単語 w_j の出現頻度とし，$v_{ij} = f_{ij}$，すなわち

$$d_i = (f_{i1}, f_{i2}, \cdots, f_{iW})^T \tag{3}$$

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，f_{w_i} を全ての文書中の単語 w_i の頻度，f_{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，

$$F = \sum_{w_i} \sum_{d_j} f_{ij} = \sum_{d_j} f_{d_j} = \sum_{w_i} f_{w_i} \tag{4}$$

の関係があるとする．v_{ij} として相対頻度を考え，$v_{ij} = \frac{f_{ij}}{F}$ とする方法や，文書の長さによる影響を解消するために $v_{ij} = \frac{f_{ij}}{f_{d_j}}$ 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 [2]．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，$tf(d_i, w_j) = \frac{f_{ij}}{F}$ 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_{ij} は，

$$v_{ij} = tf(d_i, w_j) \cdot idf(w_j) \tag{6}$$

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[12] は，

$$v_{ij} = \frac{f_{w_i}}{F} \sum_{d_j} \frac{f_{ij}}{f_{w_i}} \log \frac{\frac{f_{ij}}{f_{w_i}}}{\frac{f_{d_j}}{F}} \tag{7}$$

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．

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

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．

3 文書分類の統計的検定モデルによる考察

本研究では，文書分類の基本的性質を調べるため，最もシンプルと考えられる確率モデルを仮定し，その挙動について述べる．

いま，各文書は 2 つのクラス C_1 と C_2 から生起するものとする．一般的な仮定として，2 つのクラス C_1 と C_2 において文書と単語の出現確率が異なると考えることができる．本稿ではさらに，各文書と単語の出現確率は独立であり，確率 p_j^t をクラス C_t から出現する文書の第 j 単語 w_j の出現確率とする ($t \in \{1, 2\}$)．この時，一般性を失うことなく $j = 1, 2, \cdots, p$ については $p_j^1 > p_j^2$，$j = p+1, p+2, \cdots, p+q$ については $p_j^1 < p_j^2$，$j = p+q+1, p+q+2, \cdots, W$ については $p_j^1 = p_j^2$ であると仮定する．すなわち，初めから p 個の単語はクラス C_1 の文書で生起し易く，次の q 個の単語はクラス C_2 の文書で生起し易く，さらにそれ以降の $W-p-q$ 個の単語は両クラスで生起確率が同じであり，識別するための情報を与えない単語である．ここで，$\mathbf{p}^t$ からみた $\mathbf{p}^u$ の KL 情報量 $L(p^t; p^u)$ を

$$L(p^t; p^u) = \sum_{j=1}^{W} p_j^t \log \frac{p_j^t}{p_j^u} \tag{9}$$

と定義する．いま，ネイマン-ピアソンの定理より，2 つのクラスへの判定領域を

$$\mathcal{U}_K = \left\{ \hat{q} : \sum_{j=1}^{W} \hat{p}_j \log \frac{p_j^1}{p_j^2} \geq K \right\} \tag{10}$$

$$\mathcal{U}_K^C = \left\{ \hat{q} : \sum_{j=1}^{W} \hat{p}_j \log \frac{p_j^1}{p_j^2} < K \right\} \tag{11}$$

とかく．判別しようとしている文書の単語頻度分布が $\hat{q} \in \mathcal{U}_K$ であればクラス C_1 に分類し，$\hat{q} \in \mathcal{U}_K^C$ であればクラス C_2 に分類する．また，α をクラス C_1 から出現した文書をクラス C_2 に分類してしまう誤り（タイプ I の誤り）の確率，β をクラス C_2 から出現した文書をクラス C_1 に分類してしまう誤り（タイプ II の誤り）の確率とする．

さらに，両クラスを識別するために情報を与えない単

語を全て削除できた場合の両クラスの確率分布モデルを

$$S^t = \sum_{j=1}^{p+q} p_j^t$$

として，$\tilde{p}_j^t = p_j^t / S^t$ とおき，

$$\tilde{L}(p^t; p^u) = \sum_{j=1}^{p+q} \tilde{p}_j^t \log \frac{\tilde{p}_j^t}{\tilde{p}_j^u} \tag{12}$$

とする．このとき，仮説検定の結果として知られる Stein の補題と Sanov の定理 [3] のアナロジーとして以下の定理が得られる．

定理 1 $\beta \in (0,1)$ を固定する．α^* をタイプ II の誤りが β を超えない条件で，あらゆる決定ルールの中で最小化したタイプ I の最適誤り率とする．このとき文書長を十分長くし，$f_d \to \infty$ とすると，

$$(\alpha^*)^{1/f_d} \to \exp\left\{-S\tilde{L}(p^t; p^u)\right\} \tag{13}$$

となる．ただし，$S = S^1 = S^2$ である．

定理 2 文書がクラス C_1 から出現するとき，

$$Pr\{\hat{q} \in \mathcal{U}_K\} \leq \left\{-f_d S\tilde{L}(q^*; p^u)\right\} \tag{14}$$

で与えられる．ただし，q^* は次式で与えられる．

$$\tilde{L}(q^*; p^1) = \min_{\hat{q} \in \mathcal{U}_K} (\hat{q}; p^1) \tag{15}$$

これらの定理が意味するところを考察してみよう．定義より $S < 1$ であり，これは文書分類に意味をなす単語の出現確率を表す．文書分類問題では通常，多くの単語が自動的に切り出され，文書-単語ベクトルが自動生成されるため，分類に不要な単語も多く含まれることが考えられる．実際、通常の応用例における文書ベクトルでは、数千次元のベクトルで表現されるため、分類に関係なく、どのクラスでも同じように出現する不要単語も数多く含まれてしまう．

その不要な単語の確率が $1 - S$ であるとき，文書の判別誤りの指数部に S が現れるので，その分だけ分類精度が劣化することになる．もし，予め分類に不要な単語を除く事ができたとき，統計的検定の意味で最適なパフォーマンスが得られ，$\tilde{L}(p^t; p^u)$ がその最適な場合の指数部となる．

TF-IDF measure は，有効語に大きな重みを持たせる事で，このような不要語を排除しようとする方法であり，このような方法が有効に働くのは上で示したような誤り率の劣化によって説明できる．

4 文書分類に用いられる類似度に関する議論

ここでは，(8) 式で与えられる文書間の類似度について議論を行う．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．

前節と同じように，2 クラス問題を考え，確率 p_j^t をクラス C_t から出現する文書の第 j 単語 w_j の出現確率とする ($t \in \{1,2\}$)．これらの確率が既知であれば，

$$sim_d^*(p^t, p^u) = \sum_{j=1}^{W} p_j^t \log \frac{p_j^t}{p_j^u} \tag{16}$$

$$sim_s^*(p^t, p^u) = \sum_{j=1}^{W} p_j^t p_j^u \tag{17}$$

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．

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 $\hat{q}^t = \frac{f_{tj}}{f_{d_t}}$，$\hat{q}^u = \frac{f_{uj}}{f_{d_u}}$ を考える．ここで両者の総出現単語数は同じく n と仮定する．

$$sim_d(\hat{q}^t, \hat{q}^u) = \sum_{j=1}^{W} \hat{q}_j^t \log \frac{\hat{q}_j^t}{\hat{q}_j^u} \tag{18}$$

$$sim_s(\hat{q}^t, \hat{q}^u) = \sum_{j=1}^{W} \hat{q}_j^t \hat{q}_j^u \tag{19}$$

もし真の確率分布が分かっていれば (16) 式，(17) 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．

このとき，最尤推定量の漸近的性質から，

$$sim_d(\hat{q}^t, \hat{q}^u) = \sum_{j=1}^{W} p_j^t \log \frac{p_j^t}{p_j^u} + +O\left(\frac{1}{\sqrt{n}}\right), \quad in \quad prob. \tag{20}$$

$$sim_s(\hat{q}^t, \hat{q}^u) = \sum_{j=1}^{W} p_j^t p_j^u + O\left(\frac{1}{\sqrt{n}}\right), \quad in \quad prob. \tag{21}$$

であり，単語空間の次元数固定のもとでデータを増加させた場合，収束速度において漸近的に差異はない．

しかし，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，このような漸近論による評価はあまり現実的ではない．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．そのため前節と同様に，各

文書と単語の出現確率は独立であり，確率 p_j^t をクラス C_t から出現する文書の第 j 単語 w_j の出現確率とする ($t \in \{1,2\}$)．ここではさらに，$j = 1, 2, \cdots, p$ については $p_j^1 = r_1/p, p_j^2 = s_1/p$，$j = p+1, p+2, \cdots, p+q$ については $p_j^1 = r_2/q, p_j^2 = s_2/q$ ，$j = p+q+1, p+q+2, \cdots, W$ については $p_j^1 = p_j^2 = r/(W-p-q)$ であると仮定する．このとき，確率を知りえた場合の真の距離は，

$$sim_d^*(p^t, p^u) = pr_1 \log \frac{r_1}{s_1} + qr_2 \log \frac{r_2}{s_2} \tag{22}$$

$$sim_s^*(p^t, p^u) = \frac{\frac{r_1 s_1}{p} + \frac{r_2 s_2}{q} + \frac{r^2}{W-p-q}}{\sqrt{\frac{r_1^2}{p} + \frac{r_2^2}{q} + \frac{r^2}{W-p-q}} \sqrt{\frac{s_1^2}{p} + \frac{s_2^2}{q} + \frac{r^2}{W-p-q}}} \tag{23}$$

となる．ここで，文書分類の状況設定を表現するために，$p : q : n : W$ を一定としつつ，$W \to \infty$ という操作を考える．$pr_1 = S_1$, $qr_2 = S_2$ とし，$W \to \infty$ を考えると，高次元の空間において，各要素データの出現頻度がそれ程多くないという，相対的に少数サンプルである問題を表現できると考えられる．

$$sim_d^*(p^t, p^u) = S_1 \log \frac{r_1}{s_1} + S_2 \log \frac{r_2}{s_2} \tag{24}$$

$$sim_s^*(p^t, p^u) = S_1 s_1 + S_2 s_2 \tag{25}$$

このとき，相対的に少数サンプルの状態を保ちながら，次元数を極限に持っていく操作によっても，

$$sim_d(\hat{q}^t, \hat{q}^u) \to sim_d^*(p^t, p^u) \quad in \quad prob. \tag{26}$$

$$sim_s(\hat{q}^t, \hat{q}^u) \to sim_s^*(p^t, p^u) \quad in \quad prob. \tag{27}$$

となることがわかる．

文書分類の問題は，あるカテゴリにおいて頻出する単語群とそうでない単語群のおぼろげな括りによって表現される．通常の大きくない次元における学習問題においては，次元数は固定のもとでサンプル数の増加させたときのパフォーマンスが評価されることが多い．しかしながら，テキストマイニングの問題は，データ量は大きいが，モデルの次元数も非常に大きく，相対的にスパースの問題を想定した方法が有効となる．このような問題の特徴をうまく表現したモデルの構築と評価をさらに推し進める必要があろう．

5 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いることになる．言語データからの情報は，高次元のデータ空間に薄まって広がっており、このような問題の構造を表現するモデルの構築と分析は，実際問題においても有用な知見を与えてくれるはずである．

参考文献

[1] 河野浩之，川原 稔: "Web 検索におけるテキストマイニング", 人工知能学会誌, Vo.16, No.2, pp.212-218, (2001)

[2] 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)

[3] Richarde E. Blahut: Information Theory, Addison-Wesley Publishing Co., (1987)

[4] 伊藤哲郎: 情報検索, 昭晃堂，(1986)

[5] G. Salton and M.J.McGill: Introduction to Modern Information Retrieval, McGraw-Hill, (1983)

[6] 石岡恒憲: "記述式テストにおける自動採点システムの最新動向", 行動計量学，Vol.31, No.2, pp.67-87, (2004)

[7] 市村由美，長谷川隆明，渡部勇，佐藤光弘: "テキストマイニング－事例紹介", 人工知能学会誌, Vol.16, No.2, pp.192-200, (2001)

[8] 那須川哲哉，河野浩之，有村博紀: "テキストマイニング基盤技術", 人工知能学会誌 Vol.16, No.2, pp.201-211, (2001)

[9] 山西健司: "データ・テキストマイニングの最新動向－外れ値検出と評判分析を例に", 応用数理，Vol.12, No.4, pp.7-22, (2002)

[10] 日本語形態素解析システム「茶筌」, http://chasen.aist-nara.ac.jp/hiki/ChaSen/

[11] 北 研二: 確率的言語モデル, 東京大学出版会, (1999)

[12] A.Aizawa: "The Feature Quality: An Information Theoretic Perspective of Tfidf-like Measure", 23rd Annual International ACM SIGIR Conference, SIGIR 2000, pp.104-111, (2000)

[13] Masayuki Goto, Toshiyasu Matsushima, and Shigeichi Hirasawa: "A source model with probability distribution over word set and recurrence time theorem", IEICE Trans. on Fundamentals, Vol.E86-A, pp.2517-2525, (2003)

[14] Takashi Ishida, Masayuki Goto, Toshiyasu Matsushima, Shigeichi Hirasawa: "Properties of a word-valued source with a non-prefix-free word set", IEICE Trans. on Fundamentals, to appear, (2006)

[15] 大谷紀子: "情報検索におけるベクトル空間モデルの応用", 武蔵工業大学環境情報学部紀要，No.5, pp.99-109, (2004)

[16] 亀田雅之.: "擬似キーワードによる重要キーワードと重要文の抽出", 言語処理学会第 2 回年次大会発表論文集, pp.97-100, (1996)

[17] 伊藤潤，石田崇，後藤正幸，平澤茂一: "文間の単語共起類似度を用いた重要文抽出法", FIT 論文集, pp.83-84, (2002)

[18] H. Langseth, T. D. Nielsen. Classification using Hierarchical Naive Bayes models. Mach Learn, Vol.63, pp.135-159, Springer, (2006)

[19] 平澤 茂一, 石田 崇, 足立 鉱史, 後藤 正幸, 酒井 哲也: "文書分類技法とそのアンケート分析への応用", 経営情報学会 2005 年春季全国研究発表大会，pp.54-57, (2005)

[20] Masayuki Goto, Takashi Ishida, Shigeichi Hirasawa: "Representation method for a set of documents from the viewpoint of Bayesian statistics", 2003 IEEE International Conference on System, Man and Cybernetics, pp.4637-4642, (2003)