2004－NL－161　(12)
2004／5／14

多重トピックを表現するテキストモデルの一考察

上田 修功　斉藤 和巳
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

概要: オンライン文書量の増大に伴い，テキストモデリングは，テキスト分類，テキスト検索，テキストマイニングなどにおいて，より重要な研究課題となりつつある．文書は，しばしば，多重のトピックを持つ為，テキストのモデル化では，この多重性を考慮することが重要となる．最近，筆者らは，多重トピックテキストの確率モデルとして，パラメトリック混合モデル (PMM) を提案し，テキストの多重トピック分類問題での有効性を確認した．一方，同じ目的で，混合潜在ディリクレモデル (LDA) が独立に提案されている．本稿では，両者の文書生成モデルとしての差異について詳しく考察する．

A Study on Multi-topic Text Models

Naonori UEDA　　Kazumi SAITO
NTT Communication Science Laboratories, NTT Corporation

Abstract: As the number of online documents has been rapidly increasing, text modeling is becoming more important in information retrieval and text mining. Since a document often belongs to multiple topics, the multiplicity should be taken into account for text modeling. Recently, we proposed a novel probabilitic generative model, parametric mixture model (PMM), and showed its usefuleness for multi-topic text categorization. Independetly of our PMM, latent Dirichlet allocation (LDA) model has been proposed to model multi-topic text. In this paper, we examine the differences between these two models.

1　はじめに

近年，Web ページをはじめ膨大なオンライン文書が蓄積されつつある．それに伴い，大量文書を知識源として有効利用するためのテキストマイニング技術が渇望されている．テキストマイニング関連の応用では，主として，1 センテンスの理解というよりは，文書の大意，概念等を反映したより大まかなトピック解析が必要となる．トピックとは，スポーツ，音楽，政治といったテキストの内容を指す．文書をトピックを考慮してモデル化するテキストモデリングは，テキストマイニングの重要研究課題と言える．

テキストモデリングの代表モデルとして，ナイーブベイズ (Naive Bayes: NB) モデル [1][2] が著名である．NB モデルでは文書 d が一つのトピックについて書かれていることを前提とする．ところが，文書は二つ以上のトピックについて書かれていることが多い．例えば，Web ページの場合，ハイパーリンクにより複数のトピックから参照できる．これはそのページが多重トピックを持つことを意味する．換言すれば，テキストモデルでは一つの文書に対し，多重のトピックを取り扱えることが重要で，単一トピックを仮定する NB モデルではテキストモデルとして十分とは言えない．

潜在的意味解析 (LSA[1]) を確率モデル化した確率的潜在意味解析モデル (pLSA[2])[3] は多重の潜在トピックを仮定したモデルとして著名であるが，あくまで与えられた文書群のモデル化であり，新たな未知文書を生成するモデルではない．その意味でテキストモデルとは言えない．

既に，筆者らは多重トピックを取り扱えるテキストモデル，パラメトリック混合モデル (PMM[3]) を考案し，Web ページの多重トピック分類実験でその有効性を確認した [4][5]．一方，PMM とは独立に，多重トピックのテキストモデルとして，混合潜在ディレクレモデル (LDA[4])[6] が提案されている．本稿では，LDA と PMM における文章生成過程を詳細に比較することにより両者の差異を考察する．

2 NB モデル

LDA, PMM の土台となる単語ベースのテキスト表現および NB モデルについて説明する．

コーパス全体から抽出された単語群からストップワードの削除や語末処理などの前処理 [7] を施して得られた語彙集合を

$$\mathcal{W} = \{t_1, \ldots, t_V\} \text{ 但し, } t_i \neq t_j \ (i \neq j)$$

とする．t_i は第 i 単語 (term) を表す．V は語彙総数 (コーパスに渡る異なる単語の総数) とする．文書情報はその文書中に出現した単語 w_m を抜き出して羅列したリスト:

$$d =< w_1, \ldots, w_M > \text{ 但し, } w_m \in \mathcal{W} \qquad (1)$$

で表現できる．明らかに，単語数 M は文書ごとに異なる．同じ単語が文章中に複数回出現し得るので，$w_m,\ m = 1, \ldots, M$ の各々は必ずしも異なる単語とは限らないことに注意．以下では，このリストを文書 d と同一視し，文書 d と言えば，式 (1) のリストを指すものとする．

トピック l について書かれた文書 d の生成モデルとは，式 (1) に示した，M 単語の c を与件とする条件付き同時分布 $P(d|l) = P(w_1, \ldots, w_M|l)$ に他ならない．NB モデルでは，語の出現順序や係り受けを無視し，トピック l の下では M 単語は統計的に独立と仮定し，トピック c の文書 d の生成モデルを

$$P(d|l) = P(w_1, \ldots, w_M|l) = \prod_{m=1}^{M} P(w_m|l) \qquad (2)$$

とする．ここで，$P(w_m|l)$ はトピック l で単語 $w_m \in \mathcal{W}$ が生起する確率を表す．前述した様に，$w_1, \ldots, w_M$ には同じ単語も含まれ得るので，単語 $t_1, \ldots, t_V$ の生起確率で書き換えると，

$$P(d|l) = \prod_{m=1}^{M} \theta_{l,i}^{x_i} \qquad (3)$$

となる．ここに，$\theta_{l,i}$ は l で単語 t_i が生起する確率を，x_i は d 中に t_i が出現した回数を各々表す．明らかに，以下が成り立つ．

$$\sum_{i=1}^{V} \theta_{l,i} = 1 \quad \text{および} \quad x_1 + \cdots + x_V = M$$

式 (3) が NB モデルである．$\boldsymbol{\theta}_l = (\theta_{l,1}, \ldots, \theta_{l,V})$ は未知パラメータで，観測文書群から推定される．換言すれば，$\boldsymbol{\theta}_l$ を既知として，分布 $\theta_{l,i},\ i = 1, \ldots, V$ に従って M 単語を生成することにより文書 d が生成される．

NB モデルは単語の順番には意味を持たせていないので，生成された文書は我々が書く文書とは程遠い．しかし，このモデルで多くの文書を生成すると，それら文書群にはトピック l を特徴づける単語の頻度分布が観測され，その意味でテキストの生成モデルと言える．但し，あくまで単一トピックについてのモデル故，多重トピック文書の生成モデルではない．

3 LDA

本節では，多重トピックテキストモデルである LDA について説明する．但し，LDA は pLSA と関連が深いので，LDA の理解を容易にするために，まず，pLSA を概説し，次いで，LDA を詳述する．

3.1 pLSA

今，想定するトピック数を L とすると，pLSA における N 文書の生成過程は以下のように書ける．簡単のため，文書は全て M 単語から成るものとする．

[1] Latent Semantic Anaysis
[2] Probabilisitic Latent Semantic Analysis
[3] Parametric Mixture Model
[4] Latent Dirichlet Allocation

1. 以下を $n = 1, \ldots, N$ 回反復し N 文書を生成.

2. 以下を M 回反復し文書 d_n を生成.

2.1. 分布 $\lambda_{n,z},\ z = 1, \ldots, L$ に従って，トピック $z = l$ を選択.

2.2. 分布 $\theta_{l,i},\ i = 1, \ldots, V$ に従い，単語を 1 個生成.

上記より，d_n に単語 $t_i \in \mathcal{W}$ が出現する条件付き確率は次式となる.

$$P(t_i|n) = \sum_{l=1}^{L} \lambda_{n,l}\theta_{l,i} \tag{4}$$

d_n には，式 (4) の分布に従って単語が書かれるので，NB モデルと同様，語の独立性から，$d_n =< w_1, \ldots, w_M >$ の同時分布は次式となる.

$$P(d_n) = \prod_{i=1}^{M} \left(\sum_{l=1}^{L} \lambda_{n,l}\theta_{l,i} \right)^{x_{n,i}} \tag{5}$$

3.2 pLSA の問題点

式 (5) と式 (3) とを比較すると以下の二点で異なる.

(a) NB では，文書の生成モデルはトピック l に依存して定まるが，pLSA では全てのトピックで和がとられ，特定のトピックに依存しない.

(b) NB では，文書の生成過程が特定の文書に依存しないが，pLSA では，λ, x が文書インデックス n に依存している.

(a) より，pLSA は一つの単語を生成する毎にトピックが変り得るので，多重トピックが表現できる事を意味する．但し，(b) より，pLSA では，$\lambda_{n,l}$ を観測された N 文書から学習するため，学習データ外の第 $N+1$ 文書に対してはパラメータは未知となり，上記生成過程では新たな文書 d_{N+1} を原理的に生成できない．一方，式 (3) の NB モデルでは，対応するパラメータは文書インデックス n に依存しないので第 $N+1$ 文書を生成できる.

つまり，pLSA では，学習データを用いてモデルパラメータを推定しても，そのモデルで新たな文書を生成できないという点で，テキストモデルとして不十分であると言われている．但し，pLSA は，本来，LSA の確率モデル化を目指したもので，テキストモデリングとして提案されたわけではない．LSA は，与えられた学習データ $\mathcal{D}$ を $N \times V$ の行列 $\mathbf{D}$ と見なし，行列 $\mathbf{D}$ を特異値分解によりフロベニウスノルム[5]の観点で最良近似する手法で，pLSA は，行列 $\mathbf{D}$ の成分が，どのような確率モデルとして生成されたかを説明するモデルであり，$\mathcal{D}$ 以外の文書のモデルではない．LDA はこの問題を解決するためのモデルと位置づけられる.

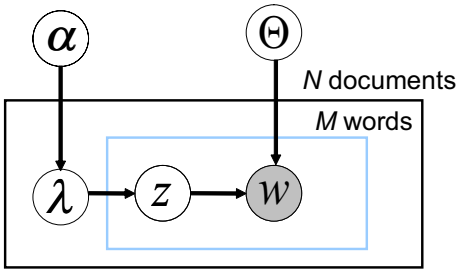

図 1: LDA の文書生成モデルのグラフィカルモデル.

3.3 LDA

LDA の文書生成過程は以下で与えられる.

1. 以下を $n = 1, \ldots, N$ 回反復し N 文書を生成.

2. 分布 Dirichlet$(\boldsymbol{\lambda}; \boldsymbol{\alpha})$ で L トピックの比率: $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)$ を定める.

3. 以下を M 回反復し一文書を生成.

3.1 分布 $\lambda_z,\ z = 1, \ldots, L$ に従って，トピック $z = l$ を選択.

3.2 分布 $\theta_{l,i},\ i = 1, \ldots, V$ に従い，単語を 1 個生成.

Dirichlet$(\boldsymbol{\lambda}; \boldsymbol{\alpha})$ はパラメータ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)$ をもつディリクレ分布で，次式で定義される.

$$\text{Dirichlet}(\boldsymbol{\lambda}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{L} \alpha_k)}{\prod_{k=1}^{L} \Gamma(\alpha_k)} \prod_{l=1}^{L} \lambda_l^{\alpha_l - 1} \tag{6}$$

ここで，$\Gamma()$ はガンマ関数である.

上記過程をグラフィカルモデルで描くと図 1 となる．グラフィカルモデルとは，変数をノードとし，変数間の依存関係を有向グラフ表現したものである．白丸のノードは潜在変数 (観測されない変数)，もしくは推定すべき未知モデルパラメータを表し，

[5] 各成分の自乗誤差の全成分に渡る総和.

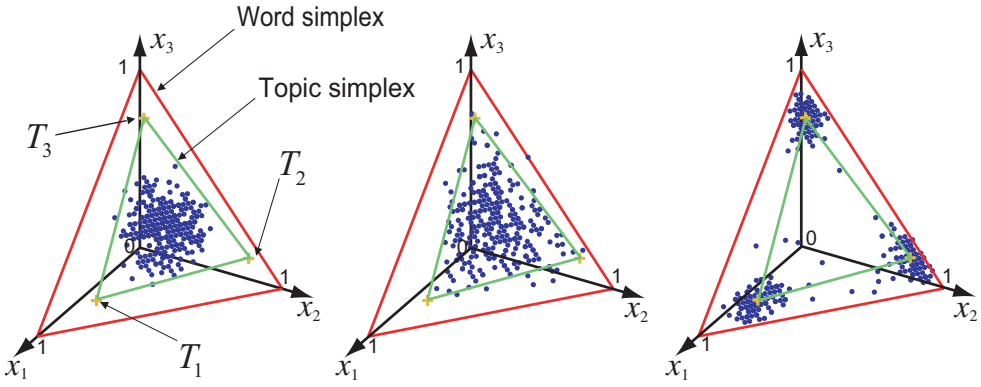

図 2: LDA において，$V = L = 3$ とし，ディリクレパラメータを変えて人工的に生成した文書の分布．各点が，一つの文書に対応する．

塗り潰された円形のノードに対応する変数は顕在変数 (既知もしくは観測される変数) を表す．矩形は反復を表す．LDA では，M 単語生成過程で潜在変数 z の値が $\boldsymbol{\lambda}$ により確率的に変化し，一つの文書中で多重トピックの単語が生成されることが分かる．また，モデルパラメータ $\Theta, \boldsymbol{\alpha}$ は文書矩形の外にあるため，新たな文書 d_{N+1} も本モデルにより生成可能であることも確認できる．

LDA による文書 d の生成モデルは次式となる．

$$P(d) = \int_{\boldsymbol{\lambda}} \mathrm{Dirichlet}(\boldsymbol{\lambda}; \boldsymbol{\alpha}) \prod_{i=1}^{V} \left(\sum_{l=1}^{L} \lambda_l \theta_{l,i} \right)^{x_i} d\boldsymbol{\lambda} \tag{7}$$

但し，上式の $\lambda_1, \ldots, \lambda_L$ に関する重積分は，$L-1$ 次元単体上 ($\lambda_l \geq 0,\ \sum_{l=1}^{L} \lambda_l = 1$) で実行されることに注意．式 (5) と式 (7) とを比較すると以下の二点で異なる．

(a) pLSA と異なり，LDA では λ が文書インデックス n に依存しない．

(b) LDA では，全てのトピック分布の取り得る全ての値 ($L-1$ 次元単体空間上に渡ってトピックの比率ベクトル $\boldsymbol{\lambda}$ が積分消去されている．但し，新たなパラメータ $\boldsymbol{\alpha}$ を持つ．

(a) は，LDA が新たな文書 d_{N+1} を生成可能なテキストモデルであることを意味する．(b) より，LDA では，式 (5) を更に $\boldsymbol{\lambda}$ に関する混合化した 2 段混合モデルという点で拡張されていることが分かる．但し，$\boldsymbol{\lambda}$ は実ベクトル故，混合化は有限混合ではなく無限混合，つまり，単体制約を満たす全ての可能な実ベクトル $\boldsymbol{\lambda}$ に関する混合となる．

LDA で生成されるトピックの多重性を確認すべく，$V = L = 3$ として LDA の生成モデルに従い人工的に，300 文書生成した結果を図 2 に示す．但し，ディリクレ分布のパラメータを三種類変えて生成した．この場合，文書は 3 種類の単語から成り，単語頻度ベクトルは 3 次元空間上の点 $\boldsymbol{x} = (x_1, x_2, x_3)$ となるが，可視化のため，$x_1 + x_2 + x_3 = 1$ にのるように正規化していることに注意．即ち，頻度ベクトルは，図 2 の各々の大きい方の正三角形で示される 2 次元単語単体 (word simplex) 上にある．また，三つのトピックに対応するパラメータベクトル $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3$ は小さい方の正三角形 (トピック単体:topic simplex) の頂点に対応する．

図 2(a) より明らかな様に，3 つのトピック T_1, T_2, T_3 のいずれにも関連する文書の分布となっており，多重性が確認できる．但し，図 2(c) のように，$\boldsymbol{\alpha}$ の各成分の値を大きくすると，混合 NB モデル

$$P(d) = \sum_{l=1}^{L} \gamma_l \prod_{i=1}^{V} \theta_{l,i}^{x_i},\ \text{但し},\ \gamma_l \geq 0,\ \sum_{l=1}^{L} \beta_l = 1 \tag{8}$$

で生成される文書分布 (但し，$V = L = 3$) と類似した分布となり，各文書の多重性が弱まり，文書群全体としては多重性を持つものの，一つの文書に注目すると単一トピックに分布していく．

4 PMM

PMM では，文書 d に対し，トピックベクトル $\boldsymbol{y} = (y_1, \ldots, y_L)$ を新たに導入する．ここに，y_l は 2 値変数で，d が第 l トピックに属す (属さない) 時，1(0) をとる．トピックベクトルは，既知あるいは未知のいずれのケースも考えられる[6]が，本稿では，LDA との比較のため，$\boldsymbol{y}$ はある潜在変数 (未知変数) として扱う．但し，$\boldsymbol{y}$ の分布 $p(\boldsymbol{y}; \beta)$ はパラメター β を持つ分布としてモデル化されているものとする．

一般に，多重トピックをもつ文書中には各トピックに関連する特徴的な単語が混合されていると考えられる．例えば，スポーツと音楽に属す文書を見ると，主としてスポーツと音楽の各々に関連する特徴語から成ることが確認できる．多重トピッククラス $\boldsymbol{y}$ において単語 t_i の生起確率 (多重トピッククラスの単語生起確率) を $\psi_i(\boldsymbol{y})$ とすると，上記知見は $\psi_i(\boldsymbol{y})$ が混合モデルとして

$$\psi_i(\boldsymbol{y}) = \sum_{l=1}^{L} h_l(\boldsymbol{y}) \theta_{l,i}, \ i = 1, \ldots, V \qquad (9)$$

と書けることを意味する．但し，混合比 $h_l(\boldsymbol{y})$ は，$y_l = 0$ となる l については $h_l(\boldsymbol{y}) = 0$ とする．

$h_l(\boldsymbol{y})$ の具体的な定義式として

$$h_l(\boldsymbol{y}) = y_l / \sum_{l'=1}^{L} y_{l'} \qquad (10)$$

が自然である [4]. 直観的には，$h_l(\boldsymbol{y})$ は文書 d が第 l トピックをもつ度合いと解釈できる．

したがって，PMM における文書生成過程は以下のようになる．

1. 以下を $n = 1, \ldots, N$ 回反復し N 文書を生成.

2. 分布 $P(\boldsymbol{y})$ でトピックベクトルを生成.

3. 以下を M 回反復し一文書を生成.

3.1 多重トピッククラス単語生起確率: $\psi_i(\boldsymbol{y}) = \sum_{l=1}^{L} h_l(\boldsymbol{y}) \theta_{l,i}$ を計算.

3.2 分布 $\psi_i(\boldsymbol{y}), \ i = 1, \ldots, V$ に従い，単語を 1 個生成.

PMM における文書分布は次式となる．

$$P(d) = \sum_{\boldsymbol{y}} P(\boldsymbol{y}; \beta) \prod_{i=1}^{V} \left(\sum_{l=1}^{L} h_l(\boldsymbol{y}) \theta_{l,i} \right)^{x_i} \qquad (11)$$

[6]テキスト分類での応用では，人手でつけたラベルを用いて教師有り学習によりモデルを学習する [4].

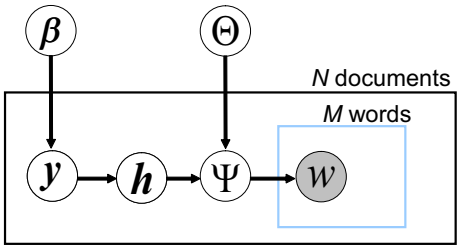

図 3: PMM の文書生成モデルのグラフィカルモデル.

5 考察

5.1 多重トピックの生成過程の違い

図 1 に示した様に，LDA ではトピックインデックスを表す潜在変数 z が単語生成矩形の内側にあることから，一文書で単語を一つ生成する際，そのたび毎に一つのトピックが” 確率的” に定められる．つまり，トピックから次から次へと遷移しながら単語が生成される．

一方，PMM では，ある文書を生成する際，多重トピッククラスの分布 $p(\boldsymbol{y})$ により多重トピッククラスが定められ，次いで，多重トピッククラスの単語生起確率 (図 3 中の Ψ) が ” 確定的” に定められ，一つの文書でその確率が固定されて M 単語が生成される．それ故，Ψ が単語生成矩形の外側に位置している．但し，Ψ は式 (9) に示した様に，既に L 個の単一トピックのパラメータ $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ の混合として定義されているので，LDA と同様，多重トピック文書が生成できる．

以上から分かるように，LDA では，単一トピックを確率的に遷移させながら NB モデルを逐次適用して多重トピック文書を生成するのに対し，PMM ではまず多重トピッククラスを作り，次いで，その多重トピッククラスをあたかも単一クラスと見なして NB モデルを適用して多重トピック文書を生成している．

5.2 文書分布の違い

この生成過程の差は文書群の分布としてどう異なるのかを図 2，図 4 を用いて説明する．図 4 は図 2 と同様の条件で PMM の文書生成過程に基づいて文書を人工的に生成した結果である．多重トピッククラスの分布 $p(\boldsymbol{y})$ は簡単のため一様分布とした．

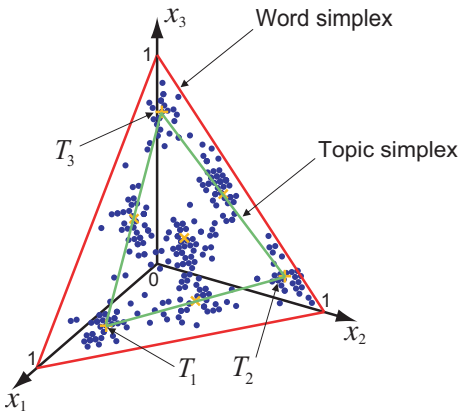

図 4: PMM での生成文書の分布．各点が，$V = L = 3$ として人工的に生成した一つの文書に対応する．

図 4 と図 2 との大きな相違点は，LDA では，文書の分布が単峰 (図 2(a),(b)) となる，もしくは，単一トピックのパラメータの周辺に分布するのに対し，PMM での文書分布は，単一トピックのパラメータの周辺，および，それらの中間的な位置にもモードを持つ点である．換言すれば，LDA では，全てのトピックをある比率で持つようにモデル化しているのに対し，PMM では部分トピックから構成される多重トピック文書をも生成できるモデル構成をとり，より” めりはり” のある多重トピック文書の生成モデルとなっている．図 4 では 7 つのモード (3 つのトピックから多重を許して 1 つ以上選択する組み合わせ数) が確認できる．

先に述べた様に，多重トピックをもつ文書は，関連単一トピックに特徴的な単語の混合からなるが，L トピックに対し，ある文書に着目した場合，必ずしもこれら全てのトピックを持つとは限らない．むしろ，その内の幾つかのトピックのみからなる事の方が多い．その意味で PMM のモデル化の方がより妥当と考えられる．また，この性質は，特に多重トピックテキスト分類において有効である [4][5].

6 まとめ

本稿では，多重トピックテキストの確率モデル LDA, PMM の文書生成過程および両者の違いについて述べた．本稿ではモデル論での議論に焦点をあて，モデルのパラメータ学習の観点での議論は紙面の都合上割愛したが，PMM は LDA に比べ，学習に要する計算時間，解の大域的最適性の観点で優れていることを追記しておく．学習の詳細は文献 [5] を参照されたい．

また，筆者らは，多重トピックテキストモデルの研究の発展形として，文書の多重トピック性を考慮した文書分類体系間の多重構造マッチング手法についても検討している．詳しくは文献 [8] を参照されたい．

参考文献

[1] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern classification (2nd ed.),” John Wiley & Sons, Inc., New York, 2000.

[2] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using EM,” Machine Learning, vol. 39, pp. 103-134. 2000.

[3] T. Hofmann, “Probabilistic latent semantic indexing,” Proc. International Conference on Information Retrieval (ACM SIGIR’99), pp. 50–57, 1999.

[4] N. Ueda and K. Saito, “Single-shot detection of multiple topics using parametric mixture models,” Proc. International Conference on Knowledge Discovery and Data Mininig (ACM SIGKDD’02), pp. 626–631, 2002.

[5] 上田修功, 斉藤和巳, “多重トピックテキストの確率モデル-パラメトリック混合モデル-,” 電子情報通信学会論文誌, vol. J87-D-II, no. 3, pp. 872–883, 2004.

[6] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” Advances in Neural Information Processign Systems (NIPS’01), 2001.

[7] C. D. Manning and H. Schtze, “Foundations of statistical natural language processing,” MIT press, Cambridge, 1999.

[8] 斉藤和巳, 上田修功, 金田有二, “確率モデルを用いた文書分類体系間の構造マッチング,” 本研究会予稿, NL-161, no.5, 2004.