多重トピックテキストの確率モデル
ーテキストモデル研究の最前線ー（2）

上田 修功 ueda@cslab.kecl.ntt.co.jp
日本電信電話(株)
NTTコミュニケーション科学基礎研究所

斉藤 和巳 saito@cslab.kecl.ntt.co.jp
日本電信電話(株)
NTTコミュニケーション科学基礎研究所

本稿の内容概要

前回，テキストモデリングではトピックの多重性をいかにモデル化するかがポイントであり，かつ，単一トピックのNBモデルを要素分布として混合した混合NBモデルでは多重性は表現できないことを説明した．これを受けて，本稿では，前回でも簡単に紹介したトピックの多重性を扱う3つのモデル（pLSA[1], LDA[2], PMM[3]）について解説する．

結論からいうと，(1) pLSAは与えられた文書集合のみを説明するモデルという点でテキストモデルとして不十分である．(2) LDA, PMMは共に多重性を扱うテキストモデルといえるが，後者の方が，モデルが単純にもかかわらず，分布のモードが複数出現する多重トピック分布を生成でき，かつ，多重トピック抽出（教師あり学習によるテキスト分類）にそのまま適用できる，さらに，モデルパラメータの学習が効率的かつ推定値の大域的最適性を理論保証するという優位性を持つ．

本稿の趣旨は上記2点の解説である．ただし，テキストモデリングでは文書生成過程の直観的理解が最重要故，本稿では，そこに全力点を置き，各モデルの学習アルゴリズムについては割愛した．学習アルゴリズムに関する詳細は参考文献を参照されたい．

以下の本文では，まず，確率モデルの理解に有用な**グラフィカルモデル**を説明し，次いで，pLSA, LDA, PMMの順に解説する．さらに，テキスト分類，テキスト検索問題といった応用問題を通して確率モデルアプローチの有用性を示す．

本稿で用いる表記を以下に整理しておく．各々の詳細説明は前回を復習していただきたい．

- 語彙集合：$\mathcal{W}=\{t_1,\ldots,t_V\}$, Vは語彙総数.
- 文書の単語リスト：$d=(w_1,\ldots,w_M)$, $w_j\in\mathcal{W}$.
- 単語頻度ベクトル：$\boldsymbol{x}=(x_1,\ldots,x_V)$
- トピック：$z\in\{1,\ldots,L\}$, Lは総トピック数.
- 単語生成確率ベクトル：$\boldsymbol{\theta}_l=(\theta_{l,1},\ldots,\theta_{l,V})$. 便宜上，$\Theta=\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_L\}$と置く．

確率モデルを図で表そう

テキストモデルの基本であるNBモデルを用いてグラフィカルモデルを説明する．**図-1**(a)にNBモデルでの文書生成過程を表現したグラフィカルモデルを示す．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．灰色に塗られた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，塗られていないノードは**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**を表す．**矩形は繰り返し**を表す．M, Nは繰り返し回数を表す．

前回解説したように，NBモデルでは，トピック$z=l$に帰属する文書を生成する際，トピック$z=l$に対応する

単語の生成確率ベクトル$\boldsymbol{\theta}_l=(\theta_{l,1},\ldots,\theta_{l,V})$に従って，語彙集合$\mathcal{W}$中の$V$種類の単語が重複を許して計$M$個生成される．つまり，NBモデルでは，単語$w$の生成にはトピックと単語生成確率に依存することから図-1(a) に示すような有向グラフとなる．

さらに，NBモデルでは，個々の文書はトピックが既知として生成されるので，トピックzは顕在変数となる．また，トピックが固定されたら，その文書中のM単語はすべて同じトピックの単語生成確率ベクトルから生成されるので，変数zは単語反復の矩形の外，かつ，文書反復の矩形の中に位置する．一方，全Lトピックの単語生成確率ベクトル集合Θは文書全体で共通故，文書反復の矩形の外に位置する．グラフィカルモデルを用いることにより，確率モデルの生成過程の直観的理解が容易になる．

次に，混合NBモデルのグラフィカルモデルを**図-1** (b)に示す．復習になるが，混合NBモデルでの文書の生成過程は以下で定義される．

—混合NBの文書生成過程—

Step 1. 確率分布$P(z)=\lambda_z, z=1,\ldots,L$に従い，1つのトピック$z=l$を選択．

Step 2. 確率分布$P(t_i|l)=\theta_{l,i}, i=1,\ldots,V$に従い，単語を$M$個を独立に生成．

以上をN回繰り返すことにより，N文書が生成される．上記生成過程についての直観的な理解が不十分な読者は前回の解説を復習していただきたい．

上記を数式表現すると次式となる．

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

$\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_L)$とすると，上記ステップのグラフィカルモデルは図-1(b)となる．NBモデルとの違いは，混合NBではトピックが確率的に定まる点である．それ故，N文書全体で見るとLトピックに渡る．ただし，前回述べたように，混合NBでは1つの文書内でのM単語の生成は，すべて同じトピックの確率分布に従うので，1つの文書は1つのトピックしか持ち得ない．式(1)にトピックに関する和が含まれているからといって，混合NBが多重トピックを表現できると考えてはいけない．再度強調しておく．混合NBの場合，1つの文書の生成過程は図-1(b)の黒線矩形の内側に示した通りで，式(1)は文書全体の確率分布を表した式であることに注意．

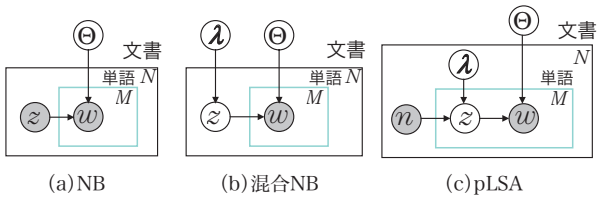

図-1 NB，混合NB，pLSAの各々における文書生成過程のグラフィカルモデル

pLSA: 多重トピックが表現可能だが…

pLSAによる文書生成過程を直観的に説明するために，今，1つの文書が1枚の原稿用紙に対応するものとする．そして，N枚の白紙の原稿用紙が与えられたとして，そこに各々M個の単語を書くことによりN文書を作成することを考える．各原稿用紙にはインデックス$n=1,\ldots,N$が付与され区別されるものとする．そして，各々の原稿用紙には，トピックを決定する確率分布$P(z|n), z=1,\ldots,L$が定義されているとする．

pLSAでは，$n=1,\ldots,N$の順に，第n原稿用紙に対し以下を行う．確率分布$P(z|n)$に従ってトピック$z=1$を決定し，次いで，確率分布$P(t_i|l)=\theta_{l,i}, i=1,\ldots,V$に従い，単語を1個生成するという過程を$M$回繰り返すことにより第$n$文書を完成させる．したがって，第$n$原稿用紙に単語$t_i\in\mathcal{W}$が出現する条件付き確率は次式となる．

$$P(t_i|n)=\sum_{z=1}^{L}P(z|n)\theta_{z,i}=\sum_{l=1}^{L}\lambda_{n,l}\theta_{l,i} \qquad (2)$$

第n原稿用紙には，式(2)の分布に従って単語が書かれるので，書かれた単語リスト$\boldsymbol{w}=(w_1,\ldots,w_M)$の同時分布は次式となる．

$$P(\boldsymbol{w}|n)=\prod_{i=1}^{M}\left(\sum_{l=1}^{L}\lambda_{n,l}\theta_{l,i}\right)^{x_{n,i}} \qquad (3)$$

式(3)では，式(1)と異なり，$\lambda_{,x}$が文書インデックスnに依存していることに注意．

pLSAのグラフィカルモデルは図-1(c)となる．図-1(b)と図-1(c)との重要な相違点を以下に整理する．

(a) 後者では，トピックzと出現単語wが単語生成過程（青線の矩形)内でペアリングされている☆1．

(b) 前者(後者)ではパラメータ$\boldsymbol{\lambda}$がN文書の黒線の矩形外

☆1 この違いは，式(1)と式(3)のxのべきの係り方の違いに反映されている．

(内) にあり，かつ，後者ではトピックzが文書インデックスnに依存して決まる．

これらの差は何を意味するのか？ (a)は，**1つの単語を生成するごとにトピックが変り得る**ことを意味する．これは式(1)と式(3)を比較しても明らかである．それ故，pLSAでは，1つの文書でも多重トピックが表現できているといえる．ならば，pLSAは多重トピックを取り扱えるテキストモデルと呼んでよいかというと，実はそうではない．その理由を以下に説明する．

トピックを決めるパラメータ$\{\lambda_{n,l}\}$はN文書からなる学習データ$\mathcal{D}$で推定されることになる．このとき，(b)より$\{\lambda_{n,l}\}$の推定値は学習データ内の第n文書に対するものであるから，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る．ということは，式(3)に示した確率モデルでは**第$N+1$文書を原理的に生成できない**．ちなみに，式(1)の混合NBモデルでは，対応するパラメータは文書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奇妙なことが起こる．上記問題はLDAの考案者らにより指摘され[2)]，この理由から，pLSAはテキストモデルとして不十分であるといわれている．

ただし，上記の批判はあくまでテキストモデリングの視点での議論であることに注意．つまり，pLSAは，そもそも潜在的意味解析（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．それ故，上記批判はpLSAの考案者Hofmannには不愉快に聞こえるであろう．つまり，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☆2，行列$\mathbf{D}$を特異値分解によりフロベニウスノルム☆3の観点で最良近似する手法がLSAであるが，pLSAは，**与えられた**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，$\mathcal{D}$以外のデータには元々興味がないのである．Hofmannとは個人的な面識はないが代弁しておく．

LDA: pLSAをテキストモデルとして拡張しよう

ここまでの内容を十分理解できた読者なら，pLSAのテキストモデルとしての問題点を解決するためには，“図-1(b)の混合NBモデルでの青線の矩形にzを含めるようにすれば，zとwとが単語生成過程でペアリングされるし，$\boldsymbol{\lambda}$も文書依存性がなくなり万事解決！”と思われるかもしれない☆4．しかし，これでは味気がない．味気がないとは…？

上記の改良では，各文書は多重トピックを持つことができ，新たな文書の生成も可能である．しかし，$\boldsymbol{\lambda}$を文書矩形の外に出すと，多重トピックの分布が全文書で共通化されるために，各文書間での多重性のパターンが似通ってしまう☆5．つまり，1つの文書でのトピックの多重性を表現できても，文書集合での多様性がなくなり，どの文書も同じような内容となり味気がなくなるのである．

そこでLDAでは，$\boldsymbol{\lambda}$を文書矩形の中に残すが，pLSAのように，各文書に依存させるのではなく，確率分布$P(\boldsymbol{\lambda})$に従って決まるものとしている．つまり，$\boldsymbol{\lambda}$は，Θやαと異なり，文書矩形の中にあるので，**1文書ごとに異なるトピック分布**$P(z)$が定められるので，N文書で見ると多様な文書が生成可能となる．$\lambda_l \geq 0, \sum_{l=1}^{L}\lambda_l = 1$故，分布$P(\boldsymbol{\lambda})$は$\boldsymbol{\lambda}=(\alpha_1,\ldots,\alpha_L)$をパラメータとするディリクレ分布としている．

LDAでの文書生成過程は以下となる．

LDAの文書生成過程

Step 1. 確率分布$P(\boldsymbol{\lambda})$でL個のトピックの比率$\lambda_1,\ldots,\lambda_L$を決定．

Step 2. 以下をM回反復する．

Step 2.1. 確率分布$P(z)=\lambda_z, z=1,\ldots,L$に従って，トピック$z=l$を選択．

Step 2.2. 確率分布$P(t_i|l)=\theta_{l,i}, i=1,\ldots,V$に従い，単語を1個生成．

以上をN回繰り返すことにより，N文書が生成される．

上記過程をグラフィカルモデルで描くと**図-2**となる．数式表現すると次式となる☆6．

☆2 第n,i要素はd_n中に単語t_iが出現した回数を表す．
☆3 各成分の自乗誤差の全成分に渡る総和．
☆4 図-1(c)で，nのノードを削除し，$\boldsymbol{\lambda}$を文書矩形の外に出したものと同じ．
☆5 一方，混合NBモデルでは，1つの文書は単一トピックながら，全文書ではよりめりはりのある多重性を持ち得る．
☆6 上記$(\lambda_1,\ldots,\lambda_L)$に渡る重積分は，単体$\lambda_1+\ldots\lambda_L=1$制約上で実行されることに注意．

$$P(\boldsymbol{d})=\int_{\boldsymbol{\lambda}} P(\boldsymbol{\lambda};\boldsymbol{\alpha})\prod_{i=1}^{V}\Big(\sum_{l=1}^{L}\lambda_l\,\theta_{l,i}\Big)^{x_{n,i}}d\boldsymbol{\lambda} \qquad (4)$$

式(4)を見ると，LDAでは，式(3)をさらに$\boldsymbol{\lambda}$に関する混合化[☆7]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．明らかに，LDAは新たな文書を生成できるという点でpLSAと異なる．

LDAは確に多重トピックを取り扱えるテキストモデルといえる．しかし，モデルパラメータα, Θの学習が難解となる．誌面の都合上，詳細は省略するが，このパラメータ推定は変分近似，もしくはマルコフ連鎖モンテカルロ法により近似的に解くことになり，推定値の局所最適性，および学習時間が膨大という実用上の問題を残している．次章で説明するPMMはこれらの問題を解決するテキストモデルである．LDAとPMMの違いはPMMを説明した後で再度議論する．

PMM: パラメトリック混合による多重化

パラメトリック混合仮説

PMMもLDAと同じく任意の多重トピック文書の生成モデルであるが，LDAとの相違点は，**トピックベクトル**および**パラメトリック混合仮説**の導入にある．トピックベクトルとは，

$$\boldsymbol{y}=(y_1,\ldots,y_L),\ y_l\in\{0,1\} \qquad (5)$$

で定義され，第lトピックを有する（しない）場合，$y_l=1(0)$とする．$\boldsymbol{y}$は，既知，未知のいずれのケースも想定される．たとえば，テキスト分類では学習データは多重トピックラベルが付与されているので，$\boldsymbol{y}$は既知となる．以下では簡単のため$\boldsymbol{y}$を既知として説明するが，$\boldsymbol{y}$が未知の場合についてもモデル化できることに注意．既知の場合は，多重トピック抽出問題に直接応用できる．これについては，**“トピック抽出への応用”**で述べる．

$\boldsymbol{y}$を既知とした場合，$\boldsymbol{y}$の可能なすべての場合についてNBモデルを構成すれば，多重トピックのテキストモデルが容易に実現可能と思われるかもしれない．しかし，トピック総数Lに対し可能な多重トピックの総クラス数は2^Lオーダと膨大[☆8]になり，単純に，多重トピッククラスごとにNBモデルを構成するのは非現実的である．そこでPMMでは，以下に説明するパラメトリック混合仮説を導入し，より簡潔なモデル構成法をとる．

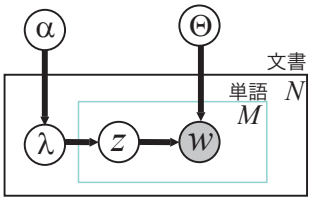

図-2　LDAにおける文書生成過程のグラフィカルモデル

一般に，多重トピックからなるテキスト中の単語は，それに関連する単一トピックに特徴的な単語の混合からなると考えられる．たとえば，“スポーツ”と“音楽”の両方のトピックに属す文書は，主に，スポーツと音楽の両者に特徴的な単語の混合からなる．これはきわめて自然であり，かつ，実際に数千のWebページでこの知見の妥当性を確認している．

この仮説に従うと，クラス$\boldsymbol{y}$の単語生成確率ベクトル$\Psi(\boldsymbol{y})=(\psi_1(\boldsymbol{y}),\ldots,\psi_V(\boldsymbol{y}))$を，その多重トピックに関連する単一トピッククラス($y_l=1$なるl)のNBモデルパラメータベクトル$\boldsymbol{\theta}_1$の線形和として次式のように表現できる．

$$\Psi(\boldsymbol{y})=\sum_{l=1}^{L}h_l(\boldsymbol{y})\theta_l \qquad (6)$$

$h_l(\boldsymbol{y})$は混合比（非負，かつ，$\sum_{l=1}^{L}h_l(\boldsymbol{y})=1$）を表し，$\boldsymbol{y}$の要素に依存し，$y_l=0$のとき，$h_l(\boldsymbol{y})=0$とする．$h_l(\boldsymbol{y})$は“第$l$トピック度”を表す量に相当する．式(6)を**パラメトリック混合仮説**と呼ぶ．

$h_l(\boldsymbol{y})$の具体的な形はいろいろ考えられるが，次式に示すように，関連する単一トピックの度合いを一様とする混合比が最も単純かつ自然である．

$$h_l(\boldsymbol{y})=\frac{y_l}{\sum_{l'=1}^{L}y_{l'}},\ l=1,\ldots,L \qquad (7)$$

たとえば，$L=3$の場合，$\Psi((1,1,0))=(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2)/2$, $\Psi((1,1,1))=(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2+\boldsymbol{\theta}_3)/3$となる．いうまでもなく，個々の多重トピック文書は，あるトピックに強い偏り（バイアス）を有することもあるが，その多重トピックに属す文書全体で平均するとそのバイアスは近似的にキャンセルされると考えられ，一様混合の式(7)は妥当といえる[☆9]．

☆7 ただし，$\boldsymbol{\lambda}$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，単体制約を満たすすべての可能な実ベクトル$\boldsymbol{\lambda}$に関する混合となる．

☆8 $L=50$だと1000兆にも及ぶ．

☆9 この妥当性についての詳細は文献3）を参照．

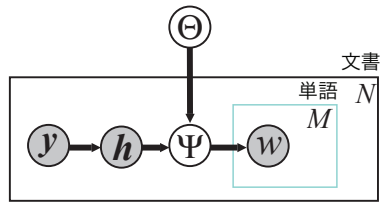

図-3　PMMにおける文書生成過程のグラフィカルモデル

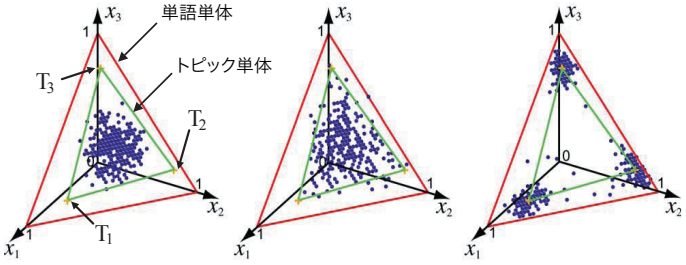

図-4　LDAのディレクトリ分布のハイパーパラメータを変えて，文書を人工的に生成した例．各点が1つの文書に対応する．

PMMでは，パラメトリック混合仮説に基づいて，$\boldsymbol{y}$に対する文書dの確率分布を次式で定義する．

$$P(d|\boldsymbol{y})=\prod_{i=1}^{V}\left(\sum_{l=1}^{L}h_l(\boldsymbol{y})\theta_{l,i}\right)^{x_i} \qquad (8)$$

PMMの文書生成過程は以下となる．

PMMの文書生成過程

Step 1. $\boldsymbol{y}$によりトピック混合比$\boldsymbol{h}=(h_1,\ldots,h_L)$を算出．

Step 2. $\boldsymbol{h}$に基づいてL個の単一トピックのNBモデルの単語生成確率ベクトル$(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_L)$を混合して$\Psi(\boldsymbol{y})=(\psi_1(\boldsymbol{y}),\ldots,\psi_V(\boldsymbol{y}))$を算出．

Step 3. 確率分布$P(t_i|l)=\psi_i(\boldsymbol{y})$, $i=1,\ldots,V$に従い，単語を独立にM個生成．

以上をN回繰り返すことにより，N文書が生成される．$\boldsymbol{y}$が既知故，Step 1，Step 2は確率的ではなく決定的に実行される．

LDAとの関係

式(4)と式(8)とを比較すると，前者でのトピック混合比はディレクレ分布から決まり，後者でのそれは$\boldsymbol{y}$から決まるという単純な違いにしか見えないかもしれない．しかし，グラフィカルモデルで比較すると，両者の単語生成過程の違いが明白となる．

図-3にPMMのグラフィカルモデルを示す．図-2と図-3の単語反復過程（青い矩形内）を比較すると分かるように，PMMでは，M回の単語生成過程において共通の確率分布Ψが用いられる．ただし，Ψ自身はすでに多重化されているという点で混合NBと異なることに注意．一方，LDAでは，1つの単語を生成するごとにトピックが確率的に決定され，そのトピックに従って単語の確率分布が決まる．つまり，LDAでは，単一トピックのNBモデルの単語生成確率ベクトル$\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_L$のいずれか1つに従って単語が生成されるのに対し，PMMでは，式(6)に示したように，$\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_L$が混合された確率ベクトル$\Psi(\boldsymbol{y})$に従って単語が生成される．それ故，PMMでは$M$反復で共通の$\Psi(\boldsymbol{y})$を用いているにもかかわらず，すでに単一トピックのNBパラメータが混合されているので多重トピックが表現可能となる．

実際どう違うのか？

LDAとPMMの文書の生成過程の違いは理解できても，両者で生成される文書は実際にどう違うのか？という素朴な疑問が沸くであろう．以下にそれについて説明する．

今，簡単のため，語彙が3個の単語 (V=3)，トピック総数が3(L=3)の場合について考える．α=(0.01 0.01 0.01), (2 2 2), (10 10 10)の3種類について，50単語(M=50)からなる文書を前述したLDAの文書生成過程に従って人工的に300個$(d_1,\ldots,d_{300})$生成した結果を**図-4**に示す．図-4中の青点が1つの文書d_nに対応する．M=50故，どの文書に対しても$x_{n,1}+x_{n,2}+x_{n,3}=50$となる．ただし，図-4では，便宜上，$M$=1となるように正規化して表示している．つまり，文書d_nに対応する正規化単語頻度ベクトル$\boldsymbol{x}$は赤い正三角形で囲まれた2次元の単体上に位置する．これを**単語単体**と呼ぶこととする．

図のT_1, T_2, T_3に示す単語生成確率ベクトル$\boldsymbol{\theta}_l=(\theta_{l,1}, \theta_{l,2}, \theta_{l,3})$, l=1,2,3も成分の和が1故，$\theta_1, \theta_2, \theta_3$は単語単体上にある．さらに，$\theta_1, \theta_2, \theta_3$を結んだ緑の三角形内の任意のベクトルもその成分の和が1故，これを**トピック単体**と呼ぶこととする．明らかに，トピック単体は単語単体上にある．これにより，単語生成確率ベクトルと単語頻度ベクトルを同一平面上に図示できる．

図-4(a)での文書の分布を見ると確かに1つの文書は

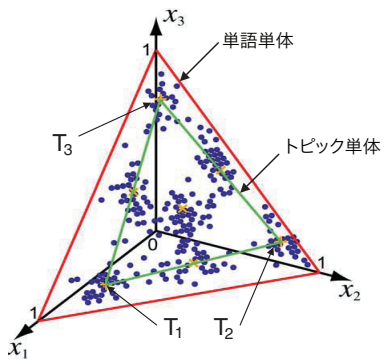

図-5　PMMにより人工的に生成された文書の分布

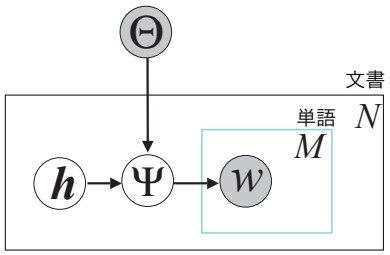

図-6　変則PMMのグラフィカルモデル

3つのトピックT_1, T_2, T_3の混合，つまり，多重トピック分布となっていることが確認できる．ただし，図-4(c)に示すパラメータに対しては，各文書は3つのトピックT_1, T_2, T_3の近傍に分布する．すなわち，αの要素の値が大きくなるにつれて，LDAの文書分布は混合NBの文書分布に近づき多重性が弱まっていく．

一方，PMMの文書生成過程での文書分布を**図-5**に示す．なお，各文書のトピックベクトル$\boldsymbol{y}$の要素は一様分布として生成している．PMMの文書分布は，遠目には7つのクラスタから構成されるように見えるであろう．実はこのクラスタの中心あたりに位置する黄色の'+'は$\Psi(\boldsymbol{y})$に対応する．つまり，L=3の場合，7通りの多重トピッククラスが考えられ[☆10]，7つのクラスタはその各々に対応している．

つまり，PMMで生成される文書分布は，T_1とT_2の多重性等がLDAに比べより明確な分布を形成できる．LDAでは多重性を有するものの，図-4(a),(b)に示すように，すべてのトピックの多重性を持った文書となるという違いがある．LDAで個々のトピックを強調すべくパラメータを変えると，今度は図-4(c)に示すような単一トピック文書となってしまう．後述するトピック抽出では，このPMMにおけるクラスタ分布の多峰性が有効利用される．

また，PMMのLDAに対する特筆すべき優位性として，**学習の効率性と最適性**が挙げられる．詳細は省略するが，PMMのモデルパラメータのMAP推定の目的関数は凸となり，逐次反復アルゴリズムでパラメータ推定が行われるものの収束した値は**大域的最適解**であることが理論保証される[3]．

$\boldsymbol{y}$が未知の場合

トピックベクトルが未知の場合のPMMも考えられる．この場合，もはや$\boldsymbol{y}$は不要で$\boldsymbol{h}$を潜在変数とするモデルとなる．このとき，$\boldsymbol{h}$とΘが決まれば，Ψが決まるので未知パラメータは$\boldsymbol{h}$とΘとなる．詳細は省略するが，このモデルの場合は，パラメータ推定での大域的最適性は成立せず，LDAと同様，局所最適解となる．

しかし，あらかじめ，図-3のモデルで$\boldsymbol{y}$既知として学習したパラメータΘを用いて，**図-6**に示すモデルで$\boldsymbol{h}$のみ未知とするとすると，$\boldsymbol{y}$が既知でΘを推定する場合と同様，$\boldsymbol{h}$の大域的最適解が得られる．この2段階モデルは**“文書検索への応用”**で説明する文書検索に応用できる．

トピック抽出への応用

ディレクトリ型のWWW検索サイトでは，あらかじめ設定されたトピック体系に基づき，人手で一般のWWWページのトピックを決めるとともに，それに応じて適切な所からリンクを張って分類している．トッピクの自動抽出は，このような作業の（半）自動化に貢献できる．我々は現実のWWWページをソフトウェアロボットで収集して実験を行った．実際に多くの人が利用している某サイトの11大分類（Arts & Humanities, Business & Economy, Computer & Internet, Education, Entertainment, Health, Recreation & Sports, Reference, Science, Social & Science, Society & Culture）を独立した

☆10 一般には，2^L-1となる．1を引いているのは，$\boldsymbol{y}$の要素がすべて零，つまり，Lトピックのいずれも持たないというケースを除外している．

	NB	SVM	kNN	NN	PMM
Arts & Humanities	41.6 (1.9)	47.1 (0.3)	40.0 (1.1)	43.3 (0.2)	**50.6** (1.0)
Business & Economy	75.0 (0.6)	74.5 (0.8)	**78.4** (0.4)	77.4 (0.5)	75.5 (0.9)
Computers & Internet	56.5 (1.3)	56.2 (1.1)	51.1 (0.8)	53.8 (1.3)	**61.0** (0.4)
Education	39.3 (1.0)	47.8 (0.8)	42.9 (0.9)	44.1 (1.0)	**51.3** (2.8)
Entertainment	54.5 (0.8)	56.9 (0.5)	47.6 (1.0)	54.9 (0.5)	**59.7** (0.4)
Health	66.4 (0.8)	**67.1** (0.3)	60.4 (0.5)	66.0 (0.4)	66.2 (0.5)
Recreation & Sports	51.8 (0.8)	52.1 (0.8)	44.4 (1.1)	49.6 (1.3)	**55.2** (0.5)
Reference	52.6 (1.1)	55.4 (0.6)	53.3 (0.5)	55.0 (1.1)	**61.1** (1.4)
Science	42.4 (0.9)	49.2 (0.7)	43.9 (0.6)	45.8 (1.3)	**51.4** (0.7)
Social & Science	41.7 (10.7)	**65.0** (1.1)	59.5 (0.9)	62.2 (2.3)	62.0 (5.1)
Society & Culture	47.2 (0.9)	51.4 (0.6)	46.4 (1.2)	50.5 (0.4)	**54.2** (0.2)

表-1　トピック抽出性能の比較

個別の問題として扱い，各問題ごとに第2階層の分類を抽出対象トピックとして用いた．収集したWWWページに対しては前処理として，各問題ごとに，冠詞等を含む571個の不要語（stop words）[4]，および，出現頻度が10以下の単語を除いて語彙集合を作成した．語彙の規模は約2万から5万単語で，各ページでの平均単語種類は百数十，トピック数は20から40，そして多重度は最大で44.4%であった．換言すれば，約60%の文書は単一トピックに属すといえる．

上記11問題の各々に対し，従来法およびPMMを適用した結果を**表-1**に示す．従来法としては，前回の解説でも説明したNBモデル，テキストマイニング分野で幅広く利用されているSVM（support vector machine）[5), 6)]やk-NN（nearest neighbour），およびNN（neural networks）を用いた．表-1は訓練データ（文書）数を2,000として学習後，訓練データにはない3,000のテストデータで評価した結果である．評価尺度としては，情報検索，テキスト分類等で共通に用いられるF尺度[4]を採用した．

F尺度は，“適合率(P)”と“再現率(R)”との調和平均で定義される．今，d_nに対する真のトピックベクトルを$\boldsymbol{y}_n=(y_{n,1},\ldots,y_{n,L})$とし，予測トピックベクトルを$\hat{y}_n=(\hat{y}_{n,1},\ldots,\hat{y}_{n,L})$とすると，適合率は予測したトピックが適合している比率: $P_n=\sum_{l=1}^{L} y_{n,l}\hat{y}_{n,l}/\sum_{l=1}^{L}\hat{y}_{n,l}$で，再現率は真のトピックが正しく予測された比率: $R_n=\sum_{l=1}^{L} y_{n,l}\hat{y}_{n,l}/\sum_{l=1}^{L} y_{n,l}$として定義される．故に，$F$尺度は$P_n, R_n$の調和平均として

$$F_n=\frac{2P_nR_n}{P_n+R_n} \qquad (9)$$

となる．実験では，訓練データとは独立した3,000ページのテストデータで$\bar{F}=\frac{1}{3000}\sum_{n=1}^{3000}F_n$として性能評価を行った．$F$尺度は単にトピックラベルベクトルの要素ごとに1/0の一致度で評価する正答率に比べ，0の一致度を評価しないという点でより厳しい評価尺度であることに注意．すなわち，過不足なく1を予測するほど，F尺度が高くなる．明らかに，F尺度の最小値（最大値）は0%（100%）となる．表中で，各問題での最良結果を太字で示し，括弧内の数値は5回の試行のF尺度の標準偏差を表す．

11種類の大半の問題で数%から約10%，PMMが従来法の最良結果を上回り，PMMの従来法に対する顕著な優位性が実証されたといえる．PMMの計算時間（2.0 Ghz pentium PCを使用）は，11問題での平均として，2,000ページの学習に約4分，3,000ページのテストデータの予測に約1分で，k-NNやNNに比べきわめて高速であった．

文書検索への応用

WWWページ検索などのように文書を探すにはキーワード検索が頻繁に利用される．これに対して，トピックに基づく文書検索ならば，キーワードがマッチしなくても概念的に近い文書群が見つかるので大変便利である．このような応用に向けて，前章で述べた11問題を用いた文書検索の実験を行った．従来法およびLDAとPMMを適用した結果を**表-2**に示す．従来法としては，テキスト間類似度として幅広く利用されているコサイン類似度に基づく手法（COS法と略記），COS法に対してIDF（inverse document frequency）を組み込んだ手法（IDF法と略記）を用いた．ここでIDFとは，単語頻度ベクトルを重み付けする処理で，テキストの分類や検索性能を一般に向上させることが知られている[4]．表-2では各問題ごとに，3,000の検索文書の各々に対し検索希望出力

	COS			IDF			LDA			PMM		
	1	10	100	1	10	100	1	10	100	1	10	100
Arts & Humanities	40.0	32.4	22.5	42.4	36.2	25.0	39.1	35.8	29.8	**45.5**	**43.5**	**37.5**
Business & Economy	71.8	68.1	62.6	73.8	69.7	62.7	72.8	70.6	67.3	**76.3**	**75.5**	**72.8**
Computers & Internet	51.1	42.9	35.1	53.2	45.4	36.1	52.3	49.3	44.0	**54.9**	**53.2**	**48.1**
Education	42.9	34.8	24.2	44.3	37.3	25.2	41.9	37.6	30.7	**45.6**	**43.2**	**37.0**
Entertainment	47.5	40.1	29.8	50.8	43.5	32.4	54.0	50.4	45.0	**57.3**	**55.1**	**49.2**
Health	58.3	51.7	40.3	60.4	54.2	42.6	62.7	59.4	52.4	**64.5**	**62.4**	**56.0**
Recreation & Sports	44.4	36.0	23.2	50.0	41.5	27.0	48.8	44.3	36.6	**54.3**	**52.0**	**44.6**
Reference	53.3	45.6	35.0	55.2	48.7	35.7	53.7	49.3	41.7	**58.6**	**56.5**	**49.7**
Science	43.8	32.9	20.6	48.4	37.9	23.2	42.7	36.3	26.4	**49.3**	**45.9**	**36.3**
Social & Science	59.4	53.9	45.1	58.9	55.1	46.2	61.5	58.1	53.0	**63.3**	**61.4**	**56.4**
Society & Culture	44.7	38.3	29.1	46.8	40.5	30.5	44.7	41.4	36.0	**50.3**	**48.8**	**42.5**
Average	50.7	43.3	33.4	53.1	46.3	35.1	52.2	48.4	42.1	**56.4**	**54.3**	**48.2**

表-2　文書検索性能の比較

件数(N=1, 10, 100)を変えながら被検索対象文書中で検索を行い，やはりF値を算出して検索性能を比較評価している．

すべてにおいて，かつ，総検索数NのすべてにおいてPMMの検索性能が従来法を顕著に上回っていることが確認できる．単語頻度ではなく，より概念的なトピックに基づく検索の優位性が実証されたといえる．N=1のときは従来法と高々約5%程度の優位性であったが，N=100では8～16%（11問題の平均では約10%）もの性能差があった．複数の類似文書を求める検索問題においてこの差は重要である．1つの検索文書に対しN=100の被検索文書を検索するのに要する時間は，平均で，COSとIDF法では0.52秒，LDAとPMMでは2.27秒であった．なお，LDAとPMMではトピック度ベクトルを求める計算時間がほとんんどで2.13秒であった．ただし，各問題ごとにPMMのパラメータの学習時間は数分であるのに対し，LDAでは丸1日以上の時間が必要であった．

また，検索されたページの内容を見ると，たとえば“自然環境問題に関する研究が重要”という趣旨のページで検索した際，“太陽電池”，“自然保護”に関するページ群が検索され，興味深いことに，これらのページ中の単語は先の“自然環境問題”のページ中には出現していなかった．つまり，COS法のようなキーワード検索では検索困難なページでも，トピックを介することにより検索可能であることを意味する．さらに，提案法の特長として，学習対象文書群を変えることにより，異なる観点で類似ページが検索できる．こうした柔軟な検索は確率モデルを用いて初めて実現できる．

あとがき

文書理解の究極は，意味理解であることはいうまでもない．意味論を真正面からとらえる言語学者から，トピックのモデル化ごときで意味理解はできないとお叱りを受けるかもしれない．実際，筆者らは統計的学習，機械学習の専門家で，言語学については素人である．しかし，近年のWebページに代表されるように，膨大に蓄積された電子テキスト群を対象とした知的情報処理が渇望される時代には，テキストモデリングのような工学的アプローチも必要だと思われる．テキストモデリング研究は意味論を無視しているのではなく，むしろ数理的に意味をモデル化しようとしているといえる．テキストモデリングの研究は国際的に見ても歴史が浅い．その意味では，この分野での世界のトップも夢ではない．読者にこの分野への参入の契機を与えることに本稿が少しでも貢献できれば幸いである．

参考文献

1）Hofmann, T.: Probabilistic Latent Semantic Indexing, Proc. International Conference on Information Retrieval (SIGIR'99), pp.50-57 (1999).
2）Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, Advances in Neural Information Processign Systems (NIPS'01) (2001).
3）Ueda, N. and Saito, K.: Single-shot Detection of Multiple Topics Using Parametric Mixture Models, Proc. International Conference on Knowledge Discovery and Data Mininig (SIGKDD'02), pp.626-631 (2002).
4）Manning, C. D. and Schtze, H.: Foundations of Statistical Natural Language Processing, MIT Press, Cambridge (1999).
5）Vapnik, V. N.: Statistical Learning Theory, John Wiley & Sons, Inc., New York (1998).
6）Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proc. the European Conference on Machine Learning (ECML'98), pp.137-142, Berlin (1998).

（平成15年11月3日受付）