Vol.2011-IFAT-101 No.7
Vol.2011-NL-200 No.7
2011/1/28

カテゴリ階層構造を考慮した確率的トピックモデルとその応用

林　幸記[†1]　江口浩二[†2,†3]　高須淳宏[†3]

高度な社会の情報化に伴い，世に生み出される情報の量は，日々加速度的に増加している．これらの大量の情報から有用な知見の抽出や，新たな知識の発見を得ることを目的とした技術が提案されてきた．なかでも，情報化社会における情報の中心を占めるテキスト形式のデータを取り扱う手法として，近年，確率的トピックモデルの有用性が注目されている．その代表的なものとして LDA（Latent Dirichlet Allocation）がよく機能することが知られている．ところで，情報の継続的な増加に伴い，大規模な情報にアクセスする有効な手段の一つとして，文書に階層カテゴリ情報を自動的に付与することによる，文書集合のインデックス化と階層化が望まれている．LDA ではカテゴリ情報を明示的にモデル化しないため，新たなモデルが求められる．そこで，本論文では，カテゴリ階層構造を持つ文書集合に適したトピックモデルとして DirTM（Directory Topic Model）を提案する．モデルパラメータをギブス・サンプリングで推定し，いくつかの実験を通して提案モデルの有効性を示す．

Probabilistic Topic Models with Category Hierarchy and its Applications

Koki Hayashi,[†1] Koji Eguchi[†2,†3] and Atsuhiro Takasu[†3]

The development towards an information society has generated a substantial volume of information. Researchers have developed techniques to extract useful findings and/or discover new knowledge from large-scale information, to date. Especially, topic modeling approaches have attracted attention recently, as means to deal with text-formatted data that play a major part in the information used in the information society. One typical method is Latent Dirichlet Allocation (LDA) that is known to work well. Meanwhile, with a growing volume of information, a way to gain access to such information is to make documents hierarchically organized by automatically assigning hierarchical categories to each document. However, LDA does not explicitly model category information, and therefore a new model needs to be developed for this objective. We propose in this paper a new topic model that we call Directory Topic Model (DirTM) to model document collections with category hierarchy. We estimate the model parameters with Gibbs sampling, and demonstrate the effectiveness of the proposed model through several experiments.

1. はじめに

高度な社会の情報化に伴い，世に生み出される情報の量は，日々加速度的に増加している．しかし，全てが有用と言えるものではなく，雑多なものも多い．これらの大量の情報を適切に取り扱い，有用な知見の抽出や，新たな知識の発見を得ることを目的とした様々な技術がこれまでに開発されてきた．なかでも，情報化社会における情報の中心を占めるテキスト形式のデータを取り扱う手法として，近年確率的トピックモデル[1)–4)] の有用性が注目されている．特に，代表的な確率的トピックモデルである潜在的ディリクレ配分法（LDA: Latent Dirichlet Allocation）[2)] は，様々なタスクへの応用やそのための拡張が研究されている．例えば，従来研究では LDA を情報検索に応用した例[5)] や，テキスト分類に適用した例[6)]，他にも種々な観点による研究[7)–9)] があり，その有用性の高さが示されてきた．基本的に LDA では，文書内のテキストデータをもとに統計的学習により潜在トピックの推定が行われる．ときに文書にはそれ自身がおおまかにどのようなテーマを扱っているかを明示的にカテゴリという形で示されている場合がある．このような文書から構成される大規模文書集合は，一般に木構造を典型とするカテゴリ構造を持っており，各カテゴリノードに関連する文書が割り当てられている．カテゴリは広い範囲の話題をカバーしており，そのカテゴリに属す文書の大まかなテーマを表わしていると言える．他にもこのカテゴリ構造から読み取れる情報は少なくなく，例えば，カテゴリ構造内ではより上位に位置するカテゴリほど抽象的であり一般的であると考えられ，同レベルのカテゴリは同程度の粒度のもの，下位のカテゴリほど特定的であると考えられる．また，各文書が属するカテゴリが上位であるほど，その文書内容も一般的な内容であるか，多様な潜在トピックを含むことが多いと考えられる．このような文書集合に対して，既存手法である LDA ではカテゴリを明示的にモデル化していないことから，潜在トピックの推定に大きな手助けとなるであろう，カテゴリ構造情報を活用していないと言える．

ところで，大規模な情報にアクセスする有効な手段の一つとして，文書に階層カテゴリ情

†1 神戸大学 大学院工学研究科 情報知能学専攻
Graduate School of Engineering, Kobe University
†2 神戸大学 大学院システム情報学研究科 情報科学専攻
Graduate School of System Informatics, Kobe University
†3 国立情報学研究所
National Institute of Informatics

Vol.2011-IFAT-101 No.7
Vol.2011-NL-200 No.7
2011/1/28

報を自動的に付与することによる，文書集合のインデックス化と階層化が望まれている．今後，情報の継続的な増加に伴い，文書を自動的に分類する技術に対する需要は増していくことが予想され，また，人手で付与されたカテゴリ階層構造を持つような文書集合の増加も見込まれる．こういった状況に反して，前述のようにLDAはカテゴリ情報を明示的にモデル化するものではなく，非階層カテゴリを用いてLDAを拡張する研究[8]があるものの，カテゴリ階層構造を活用したものは我々の知る限り存在しない．そこで，本論文では，カテゴリ階層構造を持つ文書集合に適した確率的トピックモデルとして，ディレクトリ・トピックモデル（DirTM: Directory Topic Model）を提案する．DirTMの基本的な方針は，木構造をなすカテゴリ構造における葉ノードカテゴリに属す文書のトピック多項分布の混合分布として，中間ノードカテゴリに属す文書を表現することである．この際に，各中間ノードカテゴリごとに，その子カテゴリへの辺について多項分布を仮定し，さらにその事前分布としてディリクレ分布を導入する．中間ノードから葉ノードに至るパスについてそれを構成する辺の同時確率を仮定し，中間ノードを表現するための葉ノード・トピック分布の混合比はそのパス確率に従うと仮定する．さらに，葉カテゴリごとに，文書-トピック分布のディリクレ事前分布を特定するハイパーパラメータを推定し，カテゴリごとに最適な潜在トピックが割り当てられることを目指す．

本研究では，MeSHカテゴリ構造を有し，豊富なテキストデータを持つ文書で構成されるMEDLINEコレクションに対して実際に処理を行い，推定したトピックモデルを用いて，テストデータの対数尤度を測定することで評価する．また，DirTMにより得られたトピックを素性として用いて，ロジスティック回帰モデルによる分類器を訓練し，テキスト分類実験を行うことによって，DirTMの実アプリケーションへの応用の有効性を示す．

2. 関連研究

現在までに様々なトピックモデルが考案されているが，潜在的ディリクレ配分法（LDA: Latent Dirichlet Allocation）が代表的な手法として挙げられる．本論文で提案するモデルもLDAを拡張したものであるので，まず2.1節でLDAについて説明する．次に，2.2節で他の関連するトピックモデルについて述べる．

2.1 LDA

トピックモデルとは，文書はある特徴を持った単語の分布（潜在トピック）の混合確率分布から生成されるという考えに基づいたモデルである[1)–4)]．Bleiらは，文書の潜在トピックを表す多項分布の事前分布としてディリクレ事前分布を導入した潜在的ディリクレ配分法（LDA: Latent Dirichlet Allocation）を提案した[2)]．LDAのモデル推定では文書集合における各文書の持つテキスト情報を利用し，教師なしの統計的学習を通して，文書の潜在トピック分布が推定される．

以下に，LDAによる文書の生成過程を示す[2),3)]．

(1) ディリクレ分布 $Dir(\boldsymbol{\alpha})$ から各文書 $i \in \{1, \cdots, D\}$ に関する多項分布パラメータ $\boldsymbol{\theta}_i$ を選択する．
(2) ディリクレ分布 $Dir(\boldsymbol{\beta})$ から各トピック $t \in \{1, \cdots, T\}$ に関する多項分布パラメータ $\boldsymbol{\phi}_t$ を選択する．
(3) 文書 i 内の語 w_j ($j \in \{1, \cdots, N_i\}$) それぞれに対して
  (a) 多項分布 $Mult(\boldsymbol{\theta}_i)$ からトピック z_j を選択する．
  (b) 多項分布 $Mult(\boldsymbol{\phi}_{z_j})$ から語 w_j を選択する．

ここで，$\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ はそれぞれ，文書-トピック多項分布，トピック-単語多項分布に対するディリクレ事前分布を特定するハイパーパラメータを示す．また，D は文書数，T はトピック数を表す．N_i は文書 i の文書長，w_j は文書 i の j 番目の単語を意味する．

近年，LDAに関する様々な拡張や改善を目指した研究が行われており，多くの知見が得られている．なかでも，文書-トピック多項分布に対するディリクレ事前分布のハイパーパラメータ $\boldsymbol{\alpha}$ に関しては，各トピックに対応する α_t ($t \in \{1, \cdots, T\}$) を経験的に $\alpha_t = 50/T$ とする対称ディリクレ分布が当初よく用いられ，トピック-単語多項分布に対するディリクレ事前分布のハイパーパラメータ $\boldsymbol{\beta}$ に関しては，各単語に対応する β_ω ($\omega \in \{1, \cdots, W\}$, ただし W は文書集合の語彙数を示す) を経験的に $\beta_\omega = 0.1$ とする対称ディリクレ分布がよく用いられていた[3)]．その後，固定点反復で α_t を推定することによる非対称ディリクレ分布を用いることで，モデル推定精度の改善がもたらされ，他方，$\boldsymbol{\beta}$ については固定点反復による推定は精度の劣化をもたらすため，前述の対称ディリクレ分布を用いるのがよいと報告されている[10),11)]．そこで本論文でも，文書-トピック多項分布に対するディリクレ事前分布の $\boldsymbol{\alpha}$ に関してのみ固定点反復による推定を行う．

このLDAは実際多くの場面で良く機能するが，文書のテキストデータのみを用いてモデルの推定を行っている点で，対象とする文書集合や用途によっては拡張の余地があると言える．本論文では，カテゴリ階層構造を有する文書集合に対して，より適切なトピックモデルを提案する．

2.2 その他のトピックモデル

本節では，提案するディレクトリ・トピックモデル（DirTM: Directory Topic Model）に部

Vol.2011-IFAT-101 No.7
Vol.2011-NL-200 No.7
2011/1/28

分的に関連したその他のトピックモデルについて概説する．Li らは，潜在トピックの階層的関係を推測する Pachinko allocation model（PAM）を提案した[7]．そこでは文書集合におけるテキストデータのみを用いた完全な教師なし学習が行われている．一方，DirTM ではカテゴリ階層構造が付与された文書集合に対して，階層カテゴリを補助的に用いてテキストデータの潜在トピックを推定する点で PAM とは目的が異なる．人手で付与されたカテゴリ階層構造を持つような文書集合が比較的容易に入手可能になりつつあることから，それらを利用してカテゴリが付与されていない文書を組織化することを目的としている．Blei らは，評判などの連続値またはカテゴリなどの離散値が各文書に付与された文書集合を対象に，それらを教師データとして利用しつつ潜在トピックを推定する Supervised latent Dirichlet allocation（sLDA）を提案した[8]．カテゴリ付き文書集合を対象にした LDA の拡張は，Lacoste-Julien らによる Discriminative latent Dirichlet allocation（DiscLDA）でも実現されている[9]．sLDA や DiscLDA により，各文書に非階層カテゴリが付与された文書集合を対象にして，潜在トピックを効果的に推定することができるが，いずれのモデルも階層カテゴリを想定したものではない．これに対して，DirTM ではカテゴリ階層構造を持つ文書集合を対象にしている点で着眼点が異なる．

以上のように，本論文で提案する DirTM は，カテゴリ階層構造が既知である文書集合を対象としている点，カテゴリ間の相互関係を考慮してモデル推定に活かしている点で，従来研究とは問題設定が異なると言える．

3. DirTM：ディレクトリ・トピックモデル

カテゴリが付与されている文書からなる大規模文書集合を考える．このような文書から構成される大規模文書コレクションは，一般に木構造などのカテゴリ構造を持っており，各カテゴリノードに関連する文書が割り当てられている．カテゴリは比較的広い範囲の主題をカバーし，そのカテゴリに属す文書の大まかな要約を示すと言える．つまり，同じカテゴリに属す文書では類似した事柄について記述されている可能性が高く，類似したトピック分布を持つ一方で，異なるカテゴリに属す文書とはトピック分布も異なると考えられる．また，より上位のカテゴリほど，より抽象的で，一般的な事柄について記述されていると期待され，つまりより多様な話題を含むと考えられる．また，下位のカテゴリに含まれる文書では，より特定的な事柄について記述されていると考えられる．

本論文では，このようなカテゴリ階層構造を持つ文書集合に適した確率的トピックモデルとして，ディレクトリ・トピックモデル（DirTM: Directory Topic Model）を提案する．DirTM の基本的な方針は，木構造をなすカテゴリ構造における葉ノードカテゴリに属す文書のトピック多項分布の混合分布として，中間ノードカテゴリに属す文書を表現することである．この際に，各中間ノードカテゴリごとに，その子カテゴリへの辺について多項分布を仮定し，さらにその事前分布としてディリクレ分布を導入する．中間ノードから葉ノードに至るパスについてそれを構成する辺の同時確率を仮定し，中間ノードを表現するための葉ノード・トピック分布の混合比はそのパス確率に従うと仮定する．さらに，葉カテゴリごとに，文書-トピック分布のディリクレ事前分布を特定するハイパーパラメータを推定し，カテゴリごとに最適な潜在トピックが割り当てられることを目指す．トピック-単語分布に関しては，文書の属すカテゴリによらず，対称ディリクレ事前分布を用いた．DirTM では，葉ノードカテゴリと中間ノードカテゴリにおいて，文書のモデリングが大きく異なるため，以下では，それぞれに関して個別に説明する．

3.1 葉ノードカテゴリ

DirTM では，葉ノードカテゴリに関しては，葉ノード ℓ によって異なるディリクレ分布のハイパーパラメータ $\boldsymbol{\alpha}_\ell = \{\alpha_{\ell t}\}$ $(t \in \{1, \cdots, T\})$ を用いている点を除いて，通常の LDA と同様である．LDA では，全ての文書に対して，$\boldsymbol{\alpha}$ で特定される同一のディリクレ事前分布を用いて文書-トピック多項分布パラメータをサンプリングするが，DirTM ではカテゴリごとに異なった $\boldsymbol{\alpha}_\ell$ を仮定することで，同じカテゴリに割り付けられた文書群の内容の類似性と，カテゴリごとのトピック分布の特徴の違いを，より明確にする．

DirTM において，葉ノードカテゴリ ℓ に属する文書の生成過程は以下のようになる．

(1) ディリクレ分布 $Dir(\boldsymbol{\alpha}_\ell)$ から各文書 $i \in \mathcal{D}_\ell$ に関する多項分布パラメータ $\boldsymbol{\theta}_{\ell i}$ を選択する．
(2) ディリクレ分布 $Dir(\boldsymbol{\beta})$ から各トピック $t \in \{1, \cdots, T\}$ に関する多項分布パラメータ $\boldsymbol{\phi}_t$ を選択する．
(3) 文書 i 内の語 w_j $(j \in \{1, \cdots, N_i\})$ それぞれに対して
  (a) 多項分布 $Mult(\boldsymbol{\theta}_{\ell i})$ からトピック z_j を選択する．
  (b) 多項分布 $Mult(\boldsymbol{\phi}_{z_j})$ から語 w_j を選択する．

ただし，$\mathcal{D}_\ell$ は，葉ノードカテゴリ ℓ に属する文書の集合を示す．

3.2 中間ノードカテゴリ

DirTM では，$\boldsymbol{\gamma}$ をハイパーパラメータとするディリクレ事前分布に従って，各中間ノードにおいて子ノードへの辺に関する多項分布パラメータが生成されると仮定する．ある中間ノードカテゴリ v $(v \in \{1, \cdots, V\}$ に属している文書 i を考えるとき，文書 i 内の全ての単語

Vol.2011-IFAT-101 No.7
Vol.2011-NL-200 No.7
2011/1/28

それぞれに対して，v から葉ノード $\ell \in \mathcal{L}_i$ に至るパスの集合が得られる．ここで，V は中間ノード数を示し，$\mathcal{L}_i$ は文書 i の属するカテゴリ v の下位に存在する全ての葉カテゴリの集合を指す．そして，このパスの生成確率に応じて，$\mathcal{L}_i$ 内の葉ノードカテゴリに対応するトピック分布が混合され，それが文書 i のトピック分布となる．また，そのカテゴリに属す全文書のトピック分布を均一に混合したものを，そのカテゴリのトピック分布と呼ぶ．

DirTM において，中間カテゴリ v に属する文書の生成過程は以下のようになる．

(1) ディリクレ分布 $Dir(\boldsymbol{\beta})$ から各トピック $t \in \{1, \cdots, T\}$ に関する多項分布パラメータ $\boldsymbol{\phi}_t$ を選択する．

(2) 文書 i の属する中間ノードカテゴリに対して

(a) ディリクレ分布 $Dir(\boldsymbol{\gamma})$ から各中間ノードカテゴリ v に関する多項分布パラメータ $\boldsymbol{\xi}_v$ を選択する．

(b) v が葉ノードカテゴリになるまで，以下を繰り返す．

(i) 多項分布 $Mult(\boldsymbol{\xi}_v)$ から，子ノードカテゴリ v' を選択する．

(ii) v' を v とする．

(3) 選択された葉ノードカテゴリ $\ell \in \mathcal{L}_i$ に対して

(a) ディリクレ分布 $Dir(\boldsymbol{\alpha}_\ell)$ から各文書 $i' \in \{1, \cdots, D\}$ に関する多項分布パラメータ $\boldsymbol{\theta}_{\ell i'}$ を選択する．

(b) 文書 i 内の語 w_j ($j \in \{1, \cdots, N_i\}$) それぞれに対して

(i) 葉ノードカテゴリ ℓ に属する文書の集合 $\mathcal{D}_\ell$ から一様分布に従って文書 i' を選択する．

(ii) 多項分布 $Mult(\boldsymbol{\theta}_{\ell i'})$ からトピック z_j を選択する．

(iii) 多項分布 $Mult(\boldsymbol{\phi}_{z_j})$ から語 w_j を選択する．

文書 i が属する中間ノードから葉ノード $\ell \in \mathcal{L}_i$ に至るパスに関する確率変数を $\mathbf{x}_i$ で表わすとき，DirTM のグラフィカルモデルは図 1 のようになる．

DirTM では，上記のように，中間ノードカテゴリに属する文書のトピック分布は全て，下位の葉ノードカテゴリに属する文書のトピック分布の混合分布で表現される．これにより，カテゴリ階層構造における，階層レベルによるトピックの一般性の違いを，下位のカテゴリの持つトピック参考にした上で，表現している．例えば，図 2 のようなカテゴリ階層を持つ文書集合に DirTM を適用した際の，中間ノードカテゴリ 1,2,3 の持つトピック分布について考える．葉ノードカテゴリ 4,5,6,7,8 の持つトピック分布は，ハイパーパラメータ α_ℓ をカテゴリごとに推定することにより，それぞれ特徴的で，いくつかの特定のトピックが際立っ

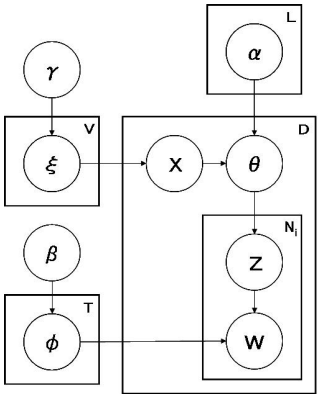

図 1 DirTM のグラフィカルモデル

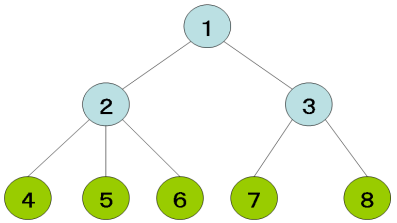

図 2 カテゴリ階層構造例

たような分布になると考えられる．これに対し，カテゴリ 1 の持つトピック分布は，カテゴリ 4,5,6,7,8, の持つトピック分布が混合されたものとなり，文書集合中に存在する全てのトピックに関連する比較的偏りの少ない分布になると考えられる．次に，カテゴリ 2 に着目すると，このカテゴリも中間ノードカテゴリであるため，特定のトピックが際立つようなトピック分布は持たないと予想される．しかし，カテゴリ 1 と違い，下位のカテゴリは 4,5,6 の三つであり，7,8 の影響を受けない．同様に，カテゴリ 3 に対しては，逆のことが言える．つまり，カテゴリ 2 と 3 では，カテゴリ 1 のように，トピックの分布の平滑化が行われるものの，下位のカテゴリの持つトピック分布の特徴もある程度反映されると考えられる．これは，階層カテゴリに属する文書の性質について考えたとき，直感的にも理解しやすく，DirTM がカテゴリ階層構造を持つ文書集合に適したモデルとなることが期待される．

3.3 DirTM の定式化

全文書の各単語に関する確率変数の集合を $\mathbf{W}$，全文書の各単語へのトピック割り当てを示す確率変数の集合を $\mathbf{Z}$ とし，全文書の各々が属する中間ノードから葉ノードに至るパスに関する確率変数の集合を $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ とするとき，DirTM のすべての確率変数と未知パラメータの同時分布は次式で表すことができる．

Vol.2011-IFAT-101 No.7
Vol.2011-NL-200 No.7
2011/1/28

$$
\begin{aligned}
p(\mathbf{W},\mathbf{X},\mathbf{Z},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\xi}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) &= p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\phi}|\boldsymbol{\beta})p(\boldsymbol{\xi}|\boldsymbol{\gamma})P(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})P(\mathbf{X}|\boldsymbol{\xi})P(\mathbf{W}|\mathbf{Z},\boldsymbol{\phi}) \\
&= \prod_{i=1}^{N}\left[\left\{\frac{\Gamma(\alpha_{\ell\Sigma})}{\prod_t\Gamma(\alpha_{\ell t})}\prod_{t=1}^{T}\theta_{\ell it}^{c(i,t)+\alpha_{\ell t}-1}\right\}\left\{\prod_{t=1}^{T}\frac{\Gamma(\beta_\Sigma)}{\prod_\omega\Gamma(\beta_\omega)}\prod_{\omega=1}^{W}\phi_{t\omega}^{c(\omega,t)+\beta_\omega-1}\right\}\right]^{\delta(u_i\in\mathcal{L})} \cdot \\
&\quad \left[\prod_{\ell\in\mathcal{L}_i}\prod_{i'\in\mathcal{D}_\ell}\frac{1}{|\mathcal{D}_\ell|}\left\{\frac{\Gamma(\alpha_{\ell\Sigma})}{\prod_t\Gamma(\alpha_{\ell t})}\prod_{t=1}^{T}\theta_{\ell i't}^{c(i',t)+\alpha_{\ell t}-1}\prod_{v=1}^{V}\frac{\Gamma(\gamma_{v\Sigma})}{\prod_u\Gamma(\gamma_{vu})}\prod_{u=1}^{U}\xi_{vu}^{\delta(v\to u\in\mathbf{x}_i)c(i',t)+\gamma_{vu}-1}\right\}\right. \\
&\quad \left.\left\{\prod_{t=1}^{T}\frac{\Gamma(\beta_\Sigma)}{\prod_\omega\Gamma(\beta_\omega)}\prod_{\omega=1}^{W}\phi_{t\omega}^{c(\omega,t)+\beta_\omega-1}\right\}\right]^{\delta(u_i\notin\mathcal{L})} \qquad (1)
\end{aligned}
$$

ここで，右辺の最初に現れる [·] は葉ノードに対応する項である．$\delta(\cdot)$ は括弧の中が満たされるときに限り 1 となるような関数である．$\mathcal{L}=\{\ell_1,\cdots,\ell_L\}$ は葉ノード集合を示し，L は葉ノード数を指す．$\alpha_{\ell t}$ ($\ell\in\{1,\cdots,L\}, t\in\{1,\cdots,T\}$) は，葉ノードカテゴリごとに異なるディリクレ分布のハイパーパラメータを示し，L は葉ノード数，T はトピックを表す．$\alpha_{\ell\Sigma}$ は全てのトピックに対する $\alpha_{\ell t}$ の総和を示す．同様に，β_Σ は全語彙にそれぞれ対する β_ω ($\omega\in\{1,\cdots,W\}$) の総和を示し，W は文書集合の語彙数を表す．$c(i,t)$ は文書 i にトピック t が割り当てられる頻度であり，$c(\omega,t)$ はトピック t に語彙 ω が割り当てられる頻度である．

また，右辺において最後に現れる [·] は中間ノードに対応する項である．$\mathcal{D}_\ell$ は葉ノードカテゴリ ℓ に属する文書の集合であり，$|\mathcal{D}_\ell|$ はその数を表す．i' は $i'\in\mathcal{D}_\ell$，すなわち，葉ノードカテゴリ ℓ に属す文書の一つを示す．$v\in\{1,\cdots,V\}$ は任意の中間ノード，$u\in\{1,\cdots,U\}$ は葉ノード・中間ノードを問わず任意のノードを示し，V は中間ノード数，U は全ノード数を示す．γ_{vu} は親子関係にあるノード対 (v,u) 間の辺 $v\to u$ に対するディリクレ事前分布のハイパーパラメータを示す．$\gamma_{v\Sigma}$ は v の全ての子ノードに対する γ_{vu} の総和を示す．

3.4 ギブス・サンプリングによるモデル推定

トピックモデルの推定には様々な手法が挙げられるが，本論文では LDA において一般的に推定精度が良いとされるギブス・サンプリング[3),4)] を用いる．推定の際に必要となる，文書 i における n 番目の単語にトピック t が割り当たる事後確率の計算式を次に示す．

$$
\begin{aligned}
&P(z_{in}=t|\mathbf{W},\mathbf{X},\mathbf{Z}_{-in},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) \\
&\propto \left[\frac{c(i,t)-1+\delta(t\neq t')+\alpha_{\ell t}}{N_i-1+\alpha_{\ell\Sigma}}\cdot\frac{c(\omega,t)-1+\delta(t\neq t')+\beta_\omega}{C_t-1+\delta(t\neq t')+\beta_\Sigma}\right]^{\delta(u_i\in\mathcal{L})} \cdot \\
&\quad \left[\prod_{\ell\in\mathcal{L}_i}\prod_{i'\in\mathcal{D}_\ell}\frac{1}{|\mathcal{D}_\ell|}\left\{\frac{c(i',t)+\alpha_{\ell t}}{N_{i'}+\alpha_{\ell\Sigma}}\prod_{v\to u\in\mathbf{x}_i}\frac{c(i',t)+\gamma_{v_{\mathbf{v}}u_{\mathbf{v}}}}{c(\mathbf{x}_i,v_{\mathbf{v}}\to\cdot)c(i',t)+\gamma_{v_{\mathbf{v}}\Sigma}}\right\}\cdot\frac{c(\omega,t)+\beta_\omega}{C_t+\beta_\Sigma}\right]^{\delta(u_i\notin\mathcal{L})} \qquad (2)
\end{aligned}
$$

これは (1) 式から導出することができる．ただし，$\mathbf{Z}_{-in}$ は $\mathbf{Z}$ から文書 i における n 番目の単語に対応する確率変数を除外したものである．また，$c(i,t)$ は，文書 i にトピック t が割り当てられる頻度であり，N_i は文書 i における全てのトピックの割り当て頻度の総和，つまり文書 i の文書長である．$c(\omega,t)$ は，トピック t に語彙 ω が割り当てられる頻度であり，C_t はトピック t に関する全ての語彙の割り当て頻度の総和である．$c(\mathbf{x}_i,v\to\cdot)$ は $\mathbf{x}_i$ が属するノード v の子ノードの数を示す．t' は文書 i における n 番目の単語に元々割り当たっていたトピックを表す．

表 1 実験に用いたコーパス

文書数	単語数	語彙数	MeSH 数
6652	391347	9019	53
葉カテゴリの持つ平均文書数	中間カテゴリの持つ平均文書数	全カテゴリの持つ平均文書数	
133.54	104.06	125.51	

4. 実　　験

提案手法の有用性を示すために，モデルの推定実験とテキスト分類実験の二つの実験を行った．

4.1 MEDLINE コレクション

MEDLINE は生物医学分野の文献データベースであり，収録されている文献数は 900 万件を超える．各文献には MeSH と呼ばれる，文献の内容を表す適切な用語が付与されており，この用語により文献を検索・管理できるようになっている．この MeSH は，階層構造を形成しており，下位にいくほど，トピックは限定的なものとなっていく．本論文ではこの MeSH をカテゴリとして用いて，実験を行う．実験で用いた文書集合は，2009 年の MEDLINE コレクションのサブセットである．具体的には，MeSH ワード algae（藻類）を根とする部分木のカテゴリ階層構造を抽出して用いた．この文書コレクションの要約を表 1 に，カテゴリ階層構造を図 3 に示す．

MEDLINE では，通常一つの文献に 10 から 15 個程度の MeSH が割り当てられているが，本論文では MeSH 階層構造の一部を抽出したため，それに属している文書のほとんどがただ一つ，多くとも 3 個程度の MeSH しか付与されていない．DirTM は各文書に対して一つのカテゴリを想定しており，MeSH を複数持つような文書は扱えないため，文書が複数の MeSH を持つ場合は階層木において最下層の MeSH を一つ残し，それ以外の MeSH を除外した．

この文書コレクションでは，情報検索システム InQuery[12)] で使用された 418 種類のストッ

Vol.2011-IFAT-101 No.7
Vol.2011-NL-200 No.7
2011/1/28

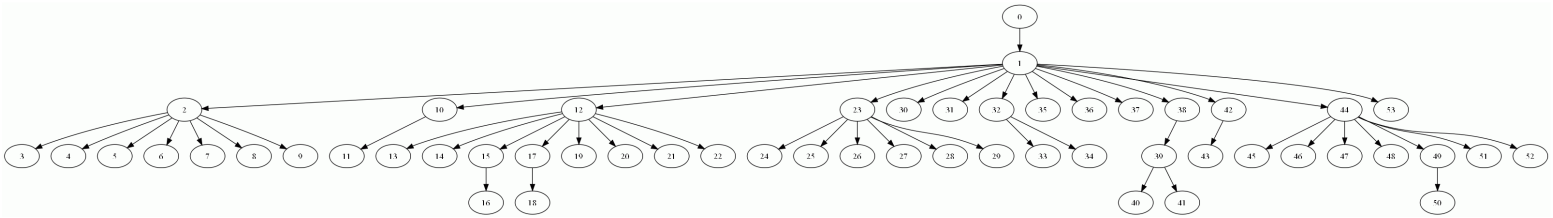

図 3 MeSH 階層構造

プワードを予め除去した．また，極端に文書長の短い文献も含まれているため，10 単語以下からなる文献を省いた．さらに，5 文書以内にしか現れない単語も排除した．後述する二つの実験ではいずれもこの文書コレクションを用いた．

4.2 テストセット対数尤度の測定実験

4.2.1 テストセット対数尤度

本実験ではモデルの評価尺度として各推定モデルに対するテストセット対数尤度を用いる．この値が大きいほど推定モデルは高精度であることを意味する．テストセット対数尤度はテストセット・パープレキシティの負の対数に等しい．テストセット・パープレキシティは言語モデルなどの統計モデルの精度を測定するためのよく知られた尺度であり[13]，トピックモデルの評価にもよく用いられる．本実験では，文書コレクションを構成する各文書から無作為に選択した単語の 90%を開発セットとし，残りの 10%をテストセットとした．

DirTM において，テストセット $\mathbf{w}^{test}$ の尤度は次式で求めることができる．

$$P(\mathbf{w}^{test}) = \prod_i \left[\prod_{\omega \in \mathbf{w}^{test}} \sum_t \frac{c(i,t)+\alpha_{\ell t}}{N_i+\alpha_{\ell\Sigma}} \cdot \frac{c(\omega,t)+\beta_\omega}{C_t+\beta_\Sigma} \right]^{\delta(u_i \in \mathcal{L})} \cdot \left[\prod_{\omega \in \mathbf{w}^{test}} \sum_t \prod_{\ell \in \mathcal{L}_i} \prod_{i' \in \mathcal{D}_\ell} \frac{1}{|\mathcal{D}_\ell|} \left\{ \frac{c(i',t)+\alpha_{\ell t}}{N_{i'}+\alpha_{\ell\Sigma}} \prod_{v \to u \in \mathbf{x}_i} \frac{c(i',t)+\gamma_{vu}}{c(\mathbf{x}_i, v \to \cdot)c(i',t)+\gamma_{v\Sigma}} \right\} \cdot \frac{c(\omega,t)+\beta_\omega}{C_t+\beta_\Sigma} \right]^{\delta(u_i \notin \mathcal{L})} \quad (3)$$

4.2.2 比較対象

本実験では，二種類の比較対象を用意した．一つは，MeSH 階層構造を全く考慮せず，文書集合全体に対して適用した，通常の LDA である．LDA は，代表的なトピックモデルであり，一つのベースラインとして重要な指標ではあるが，DirTM では，中間カテゴリのトピック分布を推定するために，葉カテゴリの文書から得られる情報を繰り返し用いているため，この点で LDA は不利な状況におかれている．そこで，比較手法の二つ目に，図 4 のようなモデルを想定した．つまり，モデル自体はカテゴリ階層構造を考慮しない通常の LDA を基本としつつ，ハイパーパラメータ $\boldsymbol{\alpha}$ をカテゴリごとに推定するモデルを用いるが，このとき，カテゴリ階層木（図中のカテゴリ 1〜5）に文書（図中の a〜e）を割り当て，中間ノードカテゴリごとにその下位のカテゴリに属する文書群を追加し，それらを中間ノードカテゴリに元々割り付けられた文書群とともに $\boldsymbol{\alpha}_\ell$ の推定のために利用した．以下では，このモデルを階層型 LDA と呼ぶことにする．この階層型 LDA と DirTM を比較することで，カテゴリ階層構造の情報を有効利用することができているのかどうか，考察を行うことができると考えた．なお，階層型 LDA では，対数尤度の測定の際に，擬似的に増やした文書 (図中の b'〜e') に関しては測定を行わず，あくまでモデル推定にのみ用いた．本実験では，以上に述べた二つを比較手法として用いた．また，後述するテキスト分類実験においても，これら二つを比較対象に用いた．

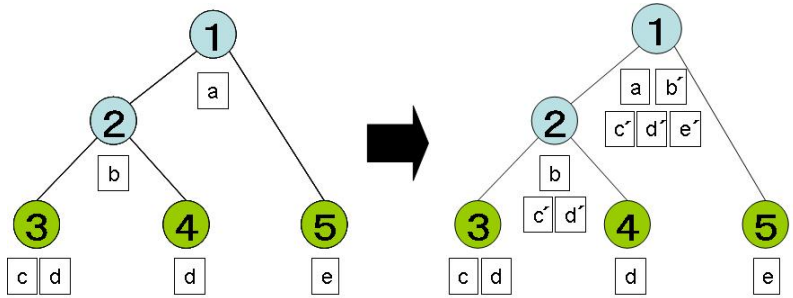

図 4 階層型 LDA

4.2.3 実験設定

本実験では，トピック数 $T = \{25, 50, 75\}$ として実験を行った．全てのモデルはギブス・サン

Vol.2011-IFAT-101 No.7
Vol.2011-NL-200 No.7
2011/1/28

プリングにより推定した．ハイパーパラメータ $\boldsymbol{\alpha}_\ell$（LDA では $\boldsymbol{\alpha}$）は，全てのモデルで，毎ギブス・スイープ終了時に，固定点反復により推定した[10),11)]．ハイパーパラメータ $\boldsymbol{\beta}$ に関しては推定を行わず，$\beta = 0.1$ に固定した[3)]．ハイパーパラメータ $\boldsymbol{\gamma}$ に関しては，$\gamma = \{0.01, 1, 100\}$ の三通りで実験を行った．

ギブス・サンプリングの収束判定条件として，推定されたモデルのテストセット対数尤度をギブス・スイープ 10 回ごとに測定し，その増加率が ±1% 以内の範囲に収まったときを収束したと見なした．

4.2.4 実験結果

実験結果を表 2〜6 にまとめた．各表では，テストセット対数尤度の単語毎平均値を「全体」として示し，さらに中間ノードカテゴリに属していた文書集合と葉ノードカテゴリに属していた文書集合についても，それぞれの単語毎平均値を示した．

表 2〜6 からわかる通り，全ての場合において，DirTM が二つの比較対象と比べて良い結果を与えた．注目すべきは，中間カテゴリの文書の対数尤度の値が，LDA，階層型 LDA の両モデルより，DirTM で特に良い値をとっている点である．つまり，DirTM がより的確にカテゴリ階層構造を捉え，より精度の高いモデル推定を実現していることがわかる．また，この実験ではハイパーパラメータ γ の違いによるテストセット対数尤度への影響は見られなかった．

4.3 テキスト分類実験

トピックモデルの適用により得られる潜在トピックを素性として用いることで，テキスト分類器の効果的な学習が実現できることが報告されている[6)]．テキスト分類器の素性として，DirTM などの各モデルによる潜在トピックを用いることにより，モデルの有用性を比較する．本論文では，テキスト分類器としてロジスティック回帰モデルを用いる．ロジスティック回帰モデルは最大エントロピー法とも呼ばれ，テキストデータのようなスパースなデータに対して特に有効であることが知られている[14)]．本実験では，ロジスティック回帰モデルによるテキスト分類実験を行うため，classias[15)] を用いた．また，訓練に際して L_2 正則化を適用した．

4.3.1 F 値

テキスト分類の実験の評価尺度として，F 値を用いた．F 値は精度と再現率の調和平均をとったもので，テキスト分類などのタスクに良く用いられる評価尺度である[14)]．

4.3.2 実験設定

文書コレクションの 90%の文書を抜き出して訓練文書セットとし，LDA，階層型 LDA，DirTM のそれぞれを用いて，事前にモデル推定を行った．また，残りの 10%をテスト文書セットとし，それらに対して 3 つのモデルそれぞれで推定されたトピック-単語分布を用いつつ，LDA のリサンプリングによって文書-トピック分布を推定した．テキスト分類器の入力として，次の三種類のデータを用いた．

- 文書中の単語およびトピック分布（LDA，階層型 LDA，DirTM）
- トピック分布のみ（LDA，階層型 LDA，DirTM）
- 文書中の出現単語のみ

このとき，単語とトピックの素性として，それぞれ文中の単語の相対頻度とトピックの割り

表 2 LDA

LDA			
トピック数	全体	中間カテゴリ	葉カテゴリ
25	-7.2874	-7.3836	-7.2672
50	-7.2027	-7.2279	-7.1978
75	-7.1725	-7.1897	-7.1693

表 3 階層型 LDA

階層型 LDA			
トピック数	全体	中間カテゴリ	葉カテゴリ
25	-7.2883	-7.3761	-7.2698
50	-7.2342	-7.2869	-7.2233
75	-7.2010	-7.2657	-7.1875

表 4 DirTM（$\gamma = 0.01$）

DirTM（$\gamma = 0.01$）			
トピック数	全体	中間カテゴリ	葉カテゴリ
25	-7.1130	-6.1488	-7.3230
50	-6.9927	-5.8640	-7.2384
75	-7.0416	-6.3886	-7.1839

表 5 DirTM（$\gamma = 1$）

DirTM（$\gamma = 1$）			
トピック数	全体	中間カテゴリ	葉カテゴリ
25	-7.1133	-6.1500	-7.3230
50	-6.9930	-5.8656	-7.2384
75	-7.0420	-6.3910	-7.1839

表 6 DirTM（$\gamma = 100$）

DirTM（$\gamma = 100$）			
トピック数	全体	中間カテゴリ	葉カテゴリ
25	-7.1142	-6.1554	-7.3230
50	-6.9938	-5.8703	-7.2384
75	-7.0428	-6.3956	-7.1839

Vol.2011-IFAT-101 No.7
Vol.2011-NL-200 No.7
2011/1/28

表 7 テキスト分類実験結果（F 値）

訓練データ	モデル	全体	中間カテゴリ	葉カテゴリ
文書中の単語 ＋トピック分布	DirTM	0.6969	0.4908	0.7919
	LDA	0.6870	0.3988	0.8208
	階層型 LDA	0.6772	0.3804	0.8150
トピック分布	DirTM	0.6772	0.4908	0.7630
	LDA	0.6752	0.3681	0.8179
	階層型 LDA	0.6693	0.3620	0.8121
文書中の単語のみ	—	0.500	0.2086	0.6358

当て回数の相対頻度を用いた．

4.3.3 実験結果

テキスト分類の実験結果を表 7 に示す．とくに中間カテゴリにおいて，DirTM の F 値が，LDA，階層型 LDA のいずれと比較しても大きく改善されていることは注意すべきである．DirTM では葉カテゴリに対する分類性能を概ね維持しつつ，中間ノードに対する分類性能を大幅に改善していることから，DirTM の狙いが達成されていると言える．全体の平均としては，DirTM で推定した潜在トピックと文書の単語情報を用いた場合が最も良い結果であったが，僅差であった．また，潜在トピックの情報だけを用いた場合でも DirTM が良い結果となった．また，いずれの場合も，文書の単語情報だけを用いた場合より，大幅な改善が見られた．

5. おわりに

本論文では，カテゴリ階層構造を有する文書集合に対し，従来までの確率的トピックモデルでは直接扱うことができなかった，テキストデータとカテゴリ階層構造の統計的な依存性を取り入れた新たなモデルを提案した．そして，モデルの推定精度を測定する実験と，テキスト分類への応用実験を行い，モデルの有効性を示した．

より多様なデータに対する詳細な評価は今後の課題である．また，大規模なデータに対する効率的なモデル推定を行うことは現実の課題として重要である．今回は応用としてテキスト分類を取り上げたが，トピックモデルの応用の範囲は広く，他のタスクへの応用を展開することも課題として挙げられる．

参考文献

1) Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proceedings of the 22nd Anuual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, USA, pp.50–57 (1999).
2) Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
3) Griffiths, T.L. and Steyvers, M.: Finding Scientific Topics, *Proceedings of the National Academy of Sciences of the United States of America*, Vol.101, pp.5228–5235 (2004).
4) Steyvers, M. and Griffiths, T.: *Handbook of Latent Semantic Analysis*, chapter21: Probabilistic Topic Models, Lawrence Erbaum Associates (2007).
5) Wei, X. and Croft, W.B.: LDA-Based Document Models for Ad-Hoc Retrieval, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, pp.178–185 (2006).
6) Phan, X.-H., Nguyen, L.-M. and Horiguchi, S.: Learning to Classify Short and Sparse Text and Web withHidden Topics from Large-scale Data Collections, *Learning to Classify Short and Sparse Text and Web with Hidden Topics from Large-scale Data Collections* (2008).
7) Li, W. and McCallum, A.: Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations, *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA (2006).
8) Blei, D.M. and McAuliffe, J.D.: Supervised Topic Models, *Advances in Neural Information Processing Systems*, Vol.20, MIT Press, pp.121–128 (2008).
9) Lacoste-Julien, S., Sha, F. and Jordan, M.I.: DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification, *Advances in Neural Information Processing Systems*, Vol.21, MIT Press, pp.897–904 (2009).
10) Minka, T.P.: Estimating a Dirichlet Distribution, Technical report, Microsoft Research, Cambridge, UK (2003).
11) Wallach, H.M., Mimno, D. and McCallum, A.: Rethinking LDA: Why Priors Matter, *Advances in Neural Information Processing Systems*, Vol.22, MIT Press, pp.1973–1981 (2010).
12) Callan, J.P., Croft, W.B. and Harding, S.M.: The INQUERY Retrieval System, *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, Valencia, Spain, pp.78–83 (1992).
13) Rabiner, L. and hwang Juang, B.: 音声認識の基礎，NTT アドバンステクノロジ株式会社 (1995).
14) Manning, C.D. and Schutze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press (1999).
15) Okazaki, N.: Classias: a collection of machine-learning algorithms for classification (2009).