アノテーション付き画像の潜在トピック階層に関するノンパラメトリックベイズモデリング

Nonparametric Bayesian modeling of latent topical hierarchies for annotated images

島廻卓史 † 江口浩二 †
Takuji SHIMAMAWARI Koji EGUCHI

1. はじめに

近年，インターネットの普及によって，世の中に存在するデータの量が爆発的に増加している．その中にはテキストデータだけでなく，画像データや映像データといったデータも数多く含まれる．特に画像データについては，写真共有コミュニティサイトである Flickr に代表されるように，画像にいくつかのアノテーション (タグ) が付与されていることがしばしばある．昨今ではアノテーション付き画像が機械学習の分野で注目されており，自動タグ付けなどの研究が盛んに行われている．

しかし，このようなデータの継続的な増加に伴い，目的のデータを探し出すことが困難になってきているのも事実で，現在さまざまなアクセス手段が必要とされている．ここで考えられる有効な手段として，潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation) [1] に代表されるトピックモデルが考えられる．トピックモデルとは，文書はある特徴を持った単語の確率分布である潜在トピックの混合分布から生成されると仮定した，文書の生成モデルである．与えられた文書から，トピックモデルを用いて潜在トピックを推定することで，情報検索，情報推薦，文書分類などへ応用できることが知られている．また，トピックモデルは拡張性が高く，テキストデータだけでなくアノテーション付き画像データなどにも適用できることが知られている．アノテーション付き画像データに適用できるトピックモデルとして，Symmetric Correspondence LDA(SymCorrLDA) [2] が挙げられる．このモデルでは画像をベクトル量子化することで画像を Bag of Visual Words と呼ばれる離散表現で表し，さらに画像に付けられたアノテーションを考慮することで，画像とそのアノテーションの相互依存性を捉えた潜在トピック生成を仮定する．

ところで，大規模データに対する有効なアクセス手段の一つとして，データに階層構造を与えるということが挙げられる．階層構造が与えられていれば，根から葉へたどっていくことで容易に探している情報を発見することができる．これを実現する代表的な階層的トピックモデルに，hierarchical LDA (hLDA) [3] がある．hLDA では文書集合に対し，潜在トピックで構成される木構造の階層を推定する．しかし hLDA をアノテーション付き画像に適用したとき，画像とそのアノテーションの相互依存性を捉えることができない．

† 神戸大学，Kobe University

そこで本論文では，SymCorrLDA を階層的トピックモデルへ拡張した，hierarchical SymCorrLDA(h-SymCorrLDA) を提案する．このモデルは，画像の特徴だけでなく，その画像に付けられたアノテーションも考慮しながら潜在トピック階層を生成する．従って，例えば犬が写った 2 枚の画像について，hLDA では犬の部分以外の画像特徴に影響されて互いに似つかないトピック分布が生成され，推定階層においても互いの画像の位置がかけ離れているかもしれない．しかし，共通のアノテーション“dog”が付与されていた場合，h-SymCorrLDA はそれを考慮することで互いに似たトピック分布を生成することができる．その結果，推定階層において，2 枚の画像は互いに近接した位置に配置されると期待される．

提案モデルの有効性を示すため，アノテーション付き画像データセットに対して，実際にテストセット対数尤度の測定実験と，正解階層と推定された階層間の正規化相互情報量の比較実験を行うことによりモデル評価を行い，既存のモデルと比較する．

2. 関連研究

まず，本研究の基礎となっている研究として，LDA とその他の関連するトピックモデルについて，その概要を説明する．

2.1 LDA

潜在的ディリクレ配分法 (LDA: Latant Dirichlet Allocation) [1] は代表的なトピックモデルである．LDA は，文書はある特徴を持った単語の分布であるトピックの混合分布から生成されると仮定する生成モデルである．また，LDA では文書-トピック多項分布とトピック-単語多項分布のそれぞれについて，ディリクレ事前分布を仮定する．

以下に LDA における文書の生成過程を示す．ここで，α，β はそれぞれ，文書-トピック多項分布およびトピック-単語多項分布に対するディリクレ事前分布のパラメータ (ハイパーパラメータ) であり，α はトピック数分，β は語彙数分の成分を持つベクトルで表される．また，θ，ϕ はそれぞれ，文書-トピック多項分布パラメータとトピック-単語多項分布パラメータである．D は文書数，T はトピック数，N_d は文書 d の文書長を表す．

（1）D 個の各文書に対して，$\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ を選択
（2）T 個の各トピックに対して，$\boldsymbol{\phi}_t \sim \mathrm{Dirichlet}(\boldsymbol{\beta})$ を選択
（3）文書 d の N_d 個の各単語 $w_{d,n}$ に対して

（a）トピック $z_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$ を選択

（b）単語 $w_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{d,n}})$ を選択

Fig 1 に LDA のグラフィカルモデルを示す．本論文では，トピックモデルの推定に周辺化ギブスサンプリング [4] を用いる．周辺化ギブスサンプリングでは，最初に各文書の各単語にトピックをランダムに割り当てる．その後，完全条件付き確率に従って各単語のトピック割り当てを更新する．これを十分回繰り返すことで，トピックを推定することができる．

LDA の周辺化ギブスサンプリングのための完全条件付き確率は以下の式で与えられる [5].

$$p(z_{d,n} = t|\mathbf{w}, \mathbf{z}_{-(d,n)}, \alpha, \beta) \propto \frac{n_{d,t}^{-(d,n)} + \alpha}{n_{d,\cdot}^{-(d,n)} + T\alpha} \cdot \frac{n_{t,w}^{-(d,n)} + \beta}{n_{t,\cdot}^{-(d,n)} + W\beta}$$

ここで，$\mathbf{w} = \{w_{d,n}\}$,$\mathbf{z} = \{z_{d,n}\}$ である．W は語彙数である．また, マイナス記号はその変数を除くことを表す．$n_{d,t}^{-(d,n)}$ と $n_{t,w}^{-(d,n)}$ はそれぞれ，d 番目の文書の n 番目の単語の割り当てを除いた，トピック t が文書 d に割り当てられた回数，トピック t が単語 w に割り当てられた回数である．また，ドット記号はその変数について総和を取っていることを表す.

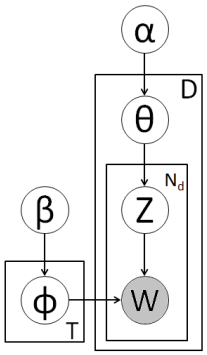

Fig. 1　A graphical model of LDA.

2.2　CorrLDA

Correspondence LDA(CorrLDA) [6] は，アノテーション付き画像や，多言語比較可能コーパスといった，複数のモードを持つデータに適用可能である．アノテーション付き画像に適用する場合，まず画像をベクトル量子化することで画像を Bag of Visual Words と呼ばれる離散表現に変換し、文書のように扱う．次に画像側かアノテーション側のどちらかについて，基準となるトピックを生成する．この基準として選んだモードをピボットモードと呼ぶ．非ピボットモードに対しては，ピボットモードで生成されたトピックを利用する.

Fig 2 に CorrLDA のグラフィカルモデルを示す．ここで，記号の右肩の (p) と (k) はそれぞれピボットモードと非ピボットモードであることを表している．$s^{(k)}$ は非ピボットモードのトピックである．CorrLDA の生成過程を以下に示す.

（1）D 個の各文書 d のピボットモード p に対して，$\boldsymbol{\theta}_d^{(p)} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}^{(p)})$ を選択

（2）T 個の各トピック t と K 個の各モード k に対して，$\boldsymbol{\phi}_t^{(k)} \sim \mathrm{Dirichlet}(\boldsymbol{\beta}^{(k)})$ を選択

（3）文書 d のピボットモード p の $N_d^{(p)}$ 個の各単語 $w_{d,n}^{(p)}$ に対して

（a）トピック $z_{d,n}^{(p)} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d^{(p)})$ を選択

（b）単語 $w_{d,n}^{(p)} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{d,n}^{(p)}}^{(p)})$ を選択

（4）文書 d の非ピボットモード $k(k \neq p)$ の $N_d^{(k)}$ 個の各単語 $w_{d,n}^{(k)}$ に対して

（a）トピック $s_{d,n}^{(k)} \sim \mathrm{Uniform}(z_1^{(p)}, \ldots, z_{N_d^{(p)}}^{(p)})$ を選択

（b）単語 $w_{d,n}^{(k)} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{s_{d,n}^{(k)}}^{(k)})$ を選択

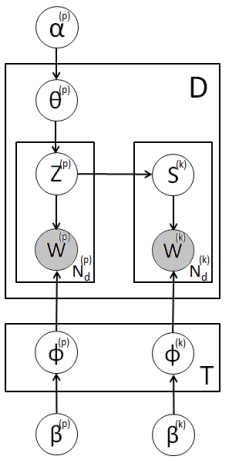

Fig. 2　A graphical model of CorrLDA.

ここでの「単語」は，テキストアノテーションにおける個々のタグと，画像の離散表現における個々の視覚単語 (Visual Words) のいずれかを示す.

CorrLDA の周辺化ギブスサンプリングのための完全条件付き確率は以下の式で与えられる.

$$\begin{aligned} p(z_{d,n}^{(p)} &= t|\mathbf{w}^{(p)}, \mathbf{z}_{-(d,n)}^{(p)}, \alpha^{(p)}, \beta^{(p)}) \\ &\propto \frac{n_{d,t}^{(p),-(d,n)} + \alpha^{(p)}}{n_{d,\cdot}^{(p),-(d,n)} + T\alpha^{(p)}} \cdot \frac{n_{t,w^{(p)}}^{(p),-(d,n)} + \beta^{(p)}}{n_{t,\cdot}^{(p),-(d,n)} + W^{(p)}\beta^{(p)}} \end{aligned}$$

$$\begin{aligned} p(s_{d,n}^{(k)} &= t|\mathbf{w}^{(k)}, \mathbf{s}_{-(d,n)}^{(k)}, \mathbf{z}^{(p)}, \beta^{(k)}) \\ &\propto \frac{n_{d,t}^{(p)}}{N_d^{(p)}} \cdot \frac{n_{t,w^{(k)}}^{(k),-(d,n)} + \beta^{(k)}}{n_{t,\cdot}^{(k),-(d,n)} + W^{(k)}\beta^{(k)}} \end{aligned}$$

このモデルはピボットモードで生成されたトピックを用いて非ピボットモードのトピックを生成するという制約により，例えばアノテーション付き画像であれば画像からそのアノテーションへの依存性を強く捉えることができる．しかし，ピボットモードをあらかじめ定めなければならないという問題がある．ピボットモードの選択は推定するモデルの質に大きく影響する．また，複数のモード間の相互依存性を捉えることができず，常にあるモードから他のモードへの一方向の依存性しか考慮されない.

2.3　SymCorrLDA

CorrLDA の問題点を改善したモデルに，Symmetric CorrLDA(SymCorrLDA) [2] がある．CorrLDA がピボットをあらかじめ定めなければならないのに対して，このモデルは最初に各単語に対してピボットを定めるフラグ (以下，ピボットフラ

グと呼ぶ) を生成し，2 つのモードから成るデータなら二項分布，3 つ以上のモードから成るデータなら多項分布によってピボットの比率を調整する．つまり SymCorrLDA は各画像における最適なピボット選択を単語レベルで推定することができる．

Fig 3 に SymCorrLDA のグラフィカルモデルを示す．ここで，π はピボットフラグ生成に関する多項分布パラメータであり，ω はそれに対するディリクレ事前分布のハイパーパラメータである．モードを K 個とした場合の SymCorrLDA の生成過程を以下に示す．

（1）D 個の各文書に対して，
  （a）K 個の各モード k に対して，$\boldsymbol{\theta}_d^{(k)} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}^{(k)})$ を選択
  （b）$\boldsymbol{\pi}_d \sim \mathrm{Dirichlet}(\omega)$ を選択
（2）T 個の各トピック t と K 個の各モード k に対して，$\boldsymbol{\phi}_t^{(k)} \sim \mathrm{Dirichlet}(\boldsymbol{\beta}^{(k)})$ を選択
（3）文書 d のモード k の $N_d^{(k)}$ 個の各単語 $w_{d,n}^{(k)}$ に対して
  （a）ピボットフラグ $x_{d,n}^{(k)} \sim \mathrm{Multinomial}(\boldsymbol{\pi}_d)$ を選択
  （b）$(x_{d,n}^{(k)} = k)$ の場合，トピック $z_{d,n}^{(k)} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d^{(k)})$ を選択
  （c）$(x_{d,n}^{(k)} = m \neq k)$ の場合，トピック $s_{d,n}^{(k)} \sim \mathrm{Uniform}(z_1^{(m)}, \ldots, z_{N_d^{(m)}}^{(m)})$ を選択
  （d）単語 $w_{d,n}^{(k)} \sim \mathrm{Multinomial}(\delta_{x_{d,n}^{(k)}=k}\boldsymbol{\phi}_{z_{d,n}^{(k)}}^{(k)} + (1 - \delta_{x_{d,n}^{(k)}=k})\boldsymbol{\phi}_{s_{d,n}^{(k)}}^{(k)})$ を選択

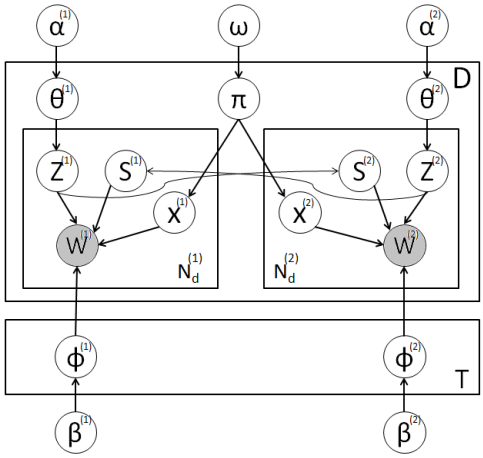

Fig. 3　A graphical model of SymCorrLDA.

モード k のピボットフラグ $x_{d,n}^{(k)} = k$ は単語 $w_{d,n}^{(k)}$ のピボットモードがそれ自身のモード k であることを示しており，$x_{d,n}^{(k)} = m$ は単語 $w_{d,n}^{(k)}$ のピボットモードがそれ自身のモード k とは異なるモード m であることを示している．ここで指示関数 δ は指定された命題が真となるとき 1，そうでなければ 0 の値をとる関数である．SymCorrLDA の周辺化ギブスサンプリングのための完全条件付き確率は以下の式で表される．

$$p(z_{d,n}^{(k)} = t, x_{d,n}^{(k)} = k | \mathbf{w}^{(k)}, \mathbf{z}_{-(d,n)}^{(k)}, \mathbf{x}_{-(d,n)}, \alpha^{(k)}, \beta^{(k)}, \omega)$$
$$\propto \frac{n_{d,k}^{-(d,n)} + \omega}{n_{d,k}^{-(d,n)} + \sum_{j \neq k} n_{d,j} + K\omega} \cdot \frac{n_{d,t}^{(k),-(d,n)} + \alpha^{(k)}}{n_{d,\cdot}^{(k),-(d,n)} + T\alpha^{(k)}} \cdot \frac{n_{t,w}^{(k),-(d,n)} + \beta^{(k)}}{n_{t,\cdot}^{(k),-(d,n)} + W^{(k)}\beta^{(k)}}$$

$$p(s_{d,n}^{(k)} = t, x_{d,n}^{(k)} = m | \mathbf{w}^{(k)}, \mathbf{s}_{-(d,n)}^{(k)}, \mathbf{z}^{(m)}, \mathbf{z}^{(k)}, \mathbf{x}_{-(d,n)}, \beta^{(k)}, \omega)$$
$$\propto \frac{n_{d,m}^{-(d,n)} + \omega}{n_{d,m}^{-(d,n)} + \sum_{j \neq m} n_{d,j} + K\omega} \cdot \frac{n_{d,t}^{(m)}}{N_d^{(m)}} \cdot \frac{n_{t,w}^{(k),-(d,n)} + \beta^{(k)}}{n_{t,\cdot}^{(k),-(d,n)} + W^{(k)}\beta^{(k)}}$$

ここで，$n_{d,k}$ と $n_{d,m}$ はアノテーション付き画像 d における単語 $w_{d,n}^{(k)}$ に対し，フラグ $x_{d,n}^{(k)} = k$ または $x_{d,n}^{(k)} = m$ が画像 d にそれぞれ割り当てられた回数である．

2.4　hLDA

hierarchical LDA(hLDA) [3] は代表的な階層的トピックモデルである．階層的トピックモデルは，あるデータに含まれる潜在トピックの階層構造をモデル化できるように，トピックモデルを拡張した生成モデルである．hLDA では，潜在トピックの階層構造が無限木 (ここでは無限個のノードで構成される有限の高さの木) を構成していると仮定する．また，各データはその部分木のうちの根ノードから葉ノードまでのあるパスに対応しており，そのパスにおける潜在トピックの混合から生成されると仮定する．木構造は nCRP(後述) によって生成される．

2.4.1　nested Chinese Restaurant Process

nested Chinese Restaurant Process(nCRP) [3] は，フラットな分布である Chinese Restaurant Process (CRP) を木構造上の分布へと拡張したものである．

まず，CRP は以下の比喩で表現することができる．あるレストランがあり，そこには無限個の順序付けられたテーブルがあるとする．また，γ は正の実数とする．最初に 1 人目の客がレストランに入店し，1 番目のテーブルに座る．n 番目 $(n \geqq 2)$ の客は，以下の式に従って k 番目のテーブルに座る．

$$p(c_d = k | c_{1:(d-1)}) = \begin{cases} \frac{m_k}{\gamma + n - 1} & (\text{if } k \text{ is existing}) \\ \frac{\gamma}{\gamma + n - 1} & (\text{if } k \text{ is existing}) \end{cases} \quad (1)$$

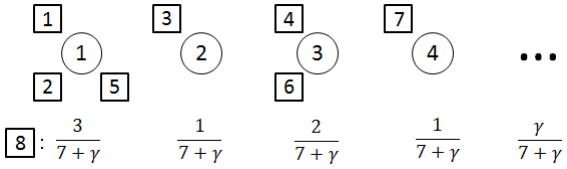

Fig. 4　A configuration of CRP.

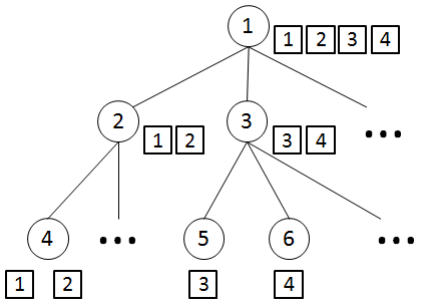

Fig. 5　A tree generated by nCRP.

ここで，m_k は k 番目のテーブルに座っている客の数である．この様子の例を Fig 4 に示す．円形はテーブルを，四角形は

データを表している．また，各テーブルの下に示されている項は，新しい客がそのテーブルに座る確率である．Fig 4 の例では，新しい客は例えば 1 番目のテーブルには確率 $\frac{3}{7+\gamma}$ で，2 番目のテーブルには確率 $\frac{1}{7+\gamma}$ で，まだ誰も座っていないテーブルには確率 $\frac{\gamma}{7+\gamma}$ で座る．式 1 からもわかるように，CRP では新しい客はすでに客が多く座っているテーブルに座りやすい．

nCRP は，客が CRP によるテーブルの選択を繰り返し行うことで，木構造を生成していく．生成される木の例を Fig 5 に示す．これは以下の比喩によって表現することができる．無限個のテーブルを持った無限個のレストランを仮定する．まず，客はあるレストランに入店し，CRP に従ってテーブルを選択する．次に，そのテーブルで指示される別のレストランに行き，再び CRP に従ってテーブルを選択する．Fig 5 の例では，根ノードが客が最初に訪れるレストランを表している．また，各ノード (レストラン) に付けられている客番号は，その客がそのレストランを訪れたことを表す．この操作を無限に行うことで，その客の木構造におけるパスを決定する．これを全ての客について行うことで，無限木中のある部分木を表現できることになる．

2.4.2 hLDA の生成過程および推定手法

Fig 6 に hLDA のグラフィカルモデルを示す．ここで，$\mathcal{T}$ は無限木，$\boldsymbol{c} = \{\boldsymbol{c}_d\}$ は各文書のパスの集合を表す．γ は nCRP のパラメータである．また，hLDA の生成過程を以下に示す．

（1）D 個の各文書に対して
- （a）$c_{d,0}$ に 無限木の根ノードを設定
- （b）各レベル $l \in \{1, \ldots, L\}$ に対して，式 (1) に従ってノード $c_{d,l}$ を選択
- （c）$\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ を選択

（2）無限木の各ノードに対応するトピック t に対して $\boldsymbol{\phi}_t \sim \mathrm{Dirichlet}(\boldsymbol{\beta})$ を選択

（3）文書 d の N_d 個の各単語 $w_{d,n}$ に対して
- （a）レベル $z_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$ を選択
- （b）単語 $w_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\phi}_{z_{d,n}})$ を選択

hLDA では，パスの推定とレベルの推定を交互に行う．パスの完全条件付き確率は以下の式で与えられる．

$$p(\mathbf{c}_d|\mathbf{z}, \mathbf{w}, \mathbf{c}_{-d}, \gamma, \beta) \propto p(\mathbf{w}_d|\mathbf{z}, \mathbf{w}_{-d}, \mathbf{c}, \beta) p(\mathbf{c}_d|\mathbf{c}_{-d}, \gamma)$$

ここで，右辺第 1 項は単語尤度であり，第 2 項は nCRP による事前分布である．単語尤度は以下の式で与えられる．

$$p(\mathbf{w}_d|\mathbf{z}, \mathbf{w}_{-d}, \mathbf{c}, \beta) = \prod_{l=1}^{L} \frac{\Gamma(n_{-d,c_{d,l},\cdot} + W\beta)}{\prod_w \Gamma(n_{-d,c_{d,l},w} + \beta)} \cdot \frac{\prod_w \Gamma(n_{d,c_{d,l},w} + n_{-d,c_{d,l},w} + \beta)}{\Gamma(n_{d,c_{d,l},\cdot} + n_{-d,c_{d,l},\cdot} + W\beta)}$$

ここで，L は木の高さである．$n_{d,c_{d,l},w}$ は文書 d について，文書 d が対応するパス中のレベル l にあたるノード $c_{d,l}$ に，語彙 w が割り当たっている回数である．また，$\Gamma(\cdot)$ はガンマ関数である．

レベルの周辺化ギブスサンプリングのための完全条件付き確率は以下の式で与えられる．

$$p(z_{d,n} = l|\mathbf{z}_{-(d,n)}, \mathbf{w}, \mathbf{c}, \alpha, \beta) \propto \frac{n_{d,l}^{-(d,n)} + \alpha}{n_{d,\cdot}^{-(d,n)} + (L+1)\alpha} \cdot \frac{n_{c_{d,l},w}^{-(d,n)} + \beta}{n_{c_{d,l},\cdot}^{-(d,n)} + W\beta}$$

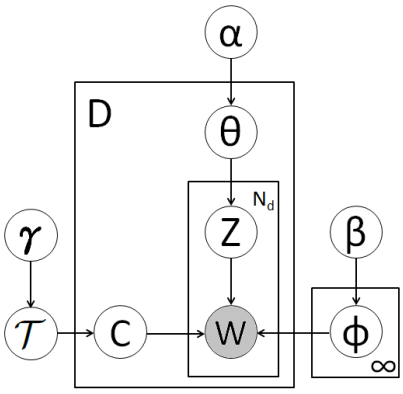

Fig. 6 A graphical model of hLDA.

3. モード間の相互依存性を考慮した階層的トピックモデル

本章では本論文で提案する h-SymCorrLDA について述べる。

3.1 h-SymCorrLDA

画像にトピックモデルを適用し，その潜在トピックに関する階層構造を推定する場合，前章で述べたように，hLDA が適用できる．しかし hLDA は画像とそのアノテーションといった，複数のモードを持つデータには対応できない．そこで，画像とそのアノテーションの相互依存関係を捉えながらトピックを生成できるモデルである SymCorrLDA を階層的トピックモデルへ拡張した，hierarchical SymCorrLDA(h-SymCorrLDA) を提案する．このモデルは，画像特徴だけでなく，その画像に付けられたアノテーションも考慮しながら潜在トピック階層を生成する．

Fig 7 に h-SymCorrLDA のグラフィカルモデルを示す．また，h-SymCorrLDA の生成過程を以下に示す．

（1）D 個の各文書に対して
- （a）$c_{d,0}$ に 無限木の根ノードを設定
- （b）各レベル $l \in \{1, \ldots, L\}$ に対して，式 (1) に従ってノード $c_{d,l}$ を選択
- （c）K 個の各モード k に対して，$\boldsymbol{\theta}_d^{(k)} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}^{(k)})$ を選択
- （d）$\boldsymbol{\pi}_d \sim \mathrm{Dirichlet}(\omega)$ を選択

（2）無限木の各ノードに対応するトピック t および K 個の各モードに対して，$\boldsymbol{\phi}_t^{(k)} \sim \mathrm{Dirichlet}(\boldsymbol{\beta}^{(k)})$ を選択

（3）文書 d のモード k の $N_d^{(k)}$ 個の各単語 $w_{d,n}^{(k)}$ に対して
- （a）ピボットフラグ $x_{d,n}^{(k)} \sim \mathrm{Multinomial}(\boldsymbol{\pi}_d)$ を選択
- （b）$(x_{d,n}^{(k)} = k)$ の場合，トピック $z_{d,n}^{(k)} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d^{(k)})$ を選択
- （c）$(x_{d,n}^{(k)} = m \neq k)$ の場合，トピック $s_{d,n}^{(k)} \sim \mathrm{Uniform}(z_1^{(m)}, \ldots, z_{N_d^{(m)}}^{(m)})$ を選択

（d）単語 $w_{d,n}^{(k)} \sim \mathrm{Multinomial}(\delta_{x_{d,n}^{(k)}=k}\boldsymbol{\phi}_{z_{d,n}^{(k)}}^{(k)} + (1 - \delta_{x_{d,n}^{(k)}=k})\boldsymbol{\phi}_{s_{d,n}^{(k)}}^{(k)})$ を選択

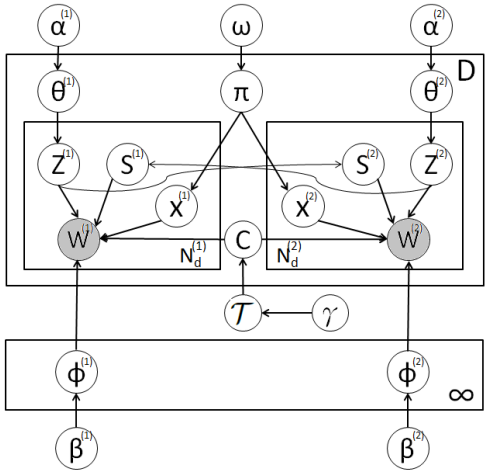

Fig. 7　A graphical model of h-SymCorrLDA.

2 つのモード ($k \in \{1, 2\}$) を仮定した h-SymCorrLDA のパスの完全条件付き確率は以下の式で与えられる．

$$p(\mathbf{c}_d|\mathbf{z}, \mathbf{s}, \mathbf{w}, \mathbf{x}, \mathbf{c}_{-d}, \gamma, \boldsymbol{\beta}, \omega) \propto p(\mathbf{w}_d^{(1)}, \mathbf{w}_d^{(2)}|\mathbf{z}, \mathbf{s}, \mathbf{w}_{-d}^{(1)}, \mathbf{w}_{-d}^{(2)}, \mathbf{x}, \mathbf{c}, \boldsymbol{\beta}, \omega) p(\mathbf{c}_d|\mathbf{c}_{-d}, \gamma)$$

ここで，$\mathbf{w} = \{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\}$，$\mathbf{z} = \{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}\}$，$\mathbf{s} = \{\mathbf{s}^{(1)}, \mathbf{s}^{(2)}\}$，$\boldsymbol{\beta} = \{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\}$ である．第 2 項は nCRP による事前分布である．第 1 項は以下の式で与えられる．

$$p(\mathbf{w}_d^{(1)}, \mathbf{w}_d^{(2)}|\mathbf{z}, \mathbf{s}, \mathbf{w}_{-d}^{(1)}, \mathbf{w}_{-d}^{(2)}, \mathbf{x}, \mathbf{c}, \boldsymbol{\beta}, \omega) = \prod_{l=1}^{L} \frac{\Gamma(n_{-d,c_{d,l},\cdot}^{(1)} + W^{(1)}\beta^{(1)})}{\prod_{w^{(1)}} \Gamma(n_{-d,c_{d,l},w^{(1)}}^{(1)} + \beta^{(1)})} \cdot \frac{\prod_{w^{(1)}} \Gamma(n_{d,c_{d,l},w^{(1)}}^{(1)} + n_{-d,c_{d,l},w^{(1)}}^{(1)} + \beta^{(1)})}{\Gamma(n_{d,c_{d,l},\cdot}^{(1)} + n_{-d,c_{d,l},\cdot}^{(1)} + W^{(1)}\beta^{(1)})} \cdot \frac{\Gamma(n_{-d,c_{d,l},\cdot}^{(2)} + W^{(2)}\beta^{(2)})}{\prod_{w^{(2)}} \Gamma(n_{-d,c_{d,l},w^{(2)}}^{(2)} + \beta^{(2)})} \cdot \frac{\prod_{w^{(2)}} \Gamma(n_{d,c_{d,l},w^{(2)}}^{(2)} + n_{-d,c_{d,l},w^{(2)}}^{(2)} + \beta^{(2)})}{\Gamma(n_{d,c_{d,l},\cdot}^{(2)} + n_{-d,c_{d,l},\cdot}^{(2)} + W^{(2)}\beta^{(2)})}$$

ここで，記号の肩の (k) は，その記号がモード k についてのものであることを表す．また，h-SymCorrLDA ではレベルは $\mathbf{z}$ としても $\mathbf{s}$ としてもサンプリングされるので，カウントについてはその両方を考慮し，双方のカウントを足し合わせたものとなっている．

周辺化ギブスサンプリングのためのレベルの完全条件付き確率は以下の式で与えられる．

$$p(z_{d,n}^{(k)} = l, x_{d,n}^{(k)} = k|\mathbf{w}^{(k)}, \mathbf{z}_{-(d,n)}^{(k)}, \mathbf{x}_{-(d,n)}, \mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega) \propto \frac{n_{d,k}^{-(d,n)} + \omega}{n_{d,k}^{-(d,n)} + \sum_{j \neq k} n_{d,j} + K\omega} \cdot \frac{n_{d,l}^{(k),-(d,n)} + \alpha^{(k)}}{n_{d,\cdot}^{(k),-(d,n)} + (L+1)\alpha^{(k)}} \cdot \frac{n_{c_{d,l},w^{(k)}}^{(k),-(d,n)} + \beta^{(k)}}{n_{c_{d,l},\cdot}^{(k),-(d,n)} + W^{(k)}\beta^{(k)}}$$

$$p(s_{d,n}^{(k)} = l, x_{d,n}^{(k)} = m|\mathbf{w}^{(k)}, \mathbf{s}_{-(d,n)}^{(k)}, \mathbf{z}^{(m)}, \mathbf{z}^{(k)}, \mathbf{x}_{-(d,n)}, \mathbf{c}, \beta^{(k)}, \omega) \propto \frac{n_{d,m}^{-(d,n)} + \omega}{n_{d,m}^{-(d,n)} + \sum_{j \neq m} n_{d,j} + K\omega} \cdot \frac{n_{d,l}^{(m)}}{N_d^{(m)}} \cdot \frac{n_{c_{d,l},w^{(k)}}^{(k),-(d,n)} + \beta^{(k)}}{n_{c_{d,l},\cdot}^{(k),-(d,n)} + W^{(k)}\beta^{(k)}}$$

3.2　h-CorrLDA

h-SymCorrLDA に似たモデルに，[7] で用いられているモデルがある．h-SymCorrLDA ではパラメータ π が各モードがピボットになる比率を調整しているが，[7] ではそのようなパラメータは導入されていない．すなわち，[7] は h-SymCorrLDA においてピボットとなるモードがあらかじめ決まっているような，特別な場合だと解釈することができる．よって，本論文では [7] で用いられているモデルを h-CorrLDA と呼ぶ．また，5. 章で示す実験で，h-CorrLDA を比較対象モデルとする．

4.　トピックモデルの画像データへの適用

文書を対象として提案されたトピックモデルを画像に適用するには，画像をベクトル量子化し，Bag of Visual Words (Bag of Features) 表現に変換する必要がある．本章では，画像の Bag of Visual Words 表現を得るまでの流れについて説明する．

4.1　Bag of Visual Words

Bag of Visual Words とは，情報検索や自然言語処理でよく使われる Bag of Words のアナロジーで，Bag of Words が文書中の単語の順序は無視して頻度のみを考えるのと同様に，Bag of Visual Words では画像における局所特徴のヒストグラムを用いて画像を表現する [8] [9]．実際の処理では，まず局所特徴点を抽出し，次にその点における局所特徴ベクトルをベクトル量子化することで，局所特徴を単語のような離散表現に置き換える．このベクトル量子化された特徴を Visual Word (視覚単語) と呼び，これが文書における単語に相当し，画像を文書のように扱うことができるようになる．以下では 4.2 節で局所特徴点の抽出について，4.3 節でベクトル量子化について説明する．

4.2　局所特徴点の抽出

局所特徴記述子には様々な種類があるが，本研究ではその中でも広く受け入れられている SIFT [10] を用いる．SIFT には画像の特徴点を検出するアルゴリズムがあり，画像中において画像特徴の変化が大きい点を特徴点として検出することができる．しかし，こうして得られた特徴点を用いるよりも，等幅の格子を仮定してその交点を特徴点と見なす方法の方が，一般画像認識などの問題において有効であることが示されている [8]．このような特徴点の抽出方法はしばしば dense sampling と呼ばれ，本研究でも dense sampling を用いて特徴点を抽出する．

4.3　ベクトル量子化

特徴点が検出できたら，SIFT を用いて各特徴点の特徴量を計算する．特徴量は 128 次元の特徴ベクトルで表される．次に，全特徴ベクトルを離散表現に置き換えるため (ベクトル量子化)，K-means アルゴリズムを用いて K 個のクラスタにクラスタリングする．このとき，各クラスタの中心点が Visual Word となる．Visual Word は K 個できるのですなわち，K-means アルゴリズムで指定する K の数が Visual Word の語彙数となる．次に，各特徴点を，その特徴点が属するクラスタのラベル (すなわちそのクラスタの Visual Word) で置き換える．以上の操作により，画像中の全ての特徴点が Visual Word に置き換わり，Bag of Visual Words 表現が得られる．

なお，ここでは慣例に従って K という記号を用いたが，これ

は 2 章及び 3 章におけるモード数を指すものではなく，$W^{(k)}$ に対応する．

5. 実　　験

提案手法である h-SymCorrLDA の有効性を示すため，モデル推定の実験を行い，いくつかの評価尺度を用いて既存モデルと比較する．

5.1 データセット

Table 1　Summary of dataset after preprocessing.

	Visual Word	Tag
Number of Images	2382	
Number of words	511818	27774
Number of vocablaries	1000	1386

MIRFLICKR Image Collection(注1) の，MIRFLICKR-25000 をデータセットとして用いる．これは，写真共有コミュニティサイトである Flickr から抽出した画像 25000 枚から成るデータセットで，各画像にはユーザーによってタグ (アノテーション) が付けられている．その平均個数は 8.94 個である．さらに，ほとんどの画像は 1 つ以上のクラスに属している．タグは，データセットにおける頻度が 20 回未満のものを除外し，全部で 1386 種類とした．クラスは 24 クラスあり，クラスに関する階層構造が与えられている．例えば，animal クラスには，dog クラスと bird クラスが子クラスとして定義されている．実験にあたり，何のクラスにも属していない画像を除外した．また，2 つ以上のクラスに属している画像については，以下の処理を行い所属クラスを 1 つに絞り込んだ．

- クラスの階層構造において最も深い位置にあるクラスを採用する．複数の同じ深さのクラスに属している場合は，データセットにおいて頻度の少ないクラスを採用する．

この処理を行った上で各クラスに属する画像のうち，付与されているタグ数の多いものから順に 100 枚ずつ選び抜き，実験に用いる画像 2382 枚を決定した．(前述のクラスに関する前処理を行った結果，baby クラスに属する画像は 82 枚となった．)

また，dense sampling において，格子幅を 30×30 ピクセルとし，局所特徴記述子のスケールを 30 ピクセルとした．なお，格子幅は 10×10 ピクセルとされることが多い [8] が，そうすると Visual Word の頻度とタグの頻度のバランスが取れないという問題が考えられる．特に h-SymCorrLDA のような，タグ側をピボットとしてトピックを生成するケースがあるモデルを推定する際は，問題となりうる．従って，格子幅を大きくして画像 1 枚あたりの Visual Word の頻度を抑え，タグとのバランスを出来るだけ保つようにした．また，ベクトル量子化のための K-means アルゴリズムにおいて，$K = 1000$ とした．これらの処理を行った上での，本実験に用いるデータセットの情報を Table 5.1 に示す．

(注1)：http://press.liacs.nl/mirflickr/

5.2 実 験 設 定

全ての階層的トピックモデルにおいて，木のレベル L は 2 とした (根ノードをレベル 0 とする)．各ハイパーパラメータについて，α は全てのモデルで 0.1 とした．全ての階層的トピックモデルについて，$\gamma = 1$ とした．また，$\omega = 1$ とした．β については，定数 C を導入し，各モード k 毎に $\beta^{(k)} = \frac{C}{W^{(k)}}$ とし，C の値を動かし実験する．これにより，各モードにおける語彙数の違いを考慮し，ハイパーパラメータに反映させることができる．

LDA と hLDA についてはモードを区別せず，各アノテーション付き画像を Visual Word とタグを区別しない単一のモードからなるデータと見なした．また，モデルのテストセット対数尤度をギブスサンプリング 10 反復毎に測定し，その変化割合が 0.1%未満に収まるとき，ギブスサンプリングを打ち切る．

5.3 テストセット対数尤度

テストセット対数尤度を測定することで，モデルの精度を評価することができる．テストセット対数尤度は以下の式で表される．

$$LL(\mathbf{w}^{test}) = \frac{\sum_d \sum_n \log P(w_{d,n}^{test} | \mathbf{w}^{train}, \mathcal{M})}{\sum_d N_d^{test}}$$

ただし，$\mathbf{w}^{test}$ はテストセット，$\mathbf{w}^{train}$ は訓練セット，$\mathcal{M}$ は評価対象のモデル，N_d^{test} はテストセットにおける文書 d の単語数を示す．テストセット対数尤度はその値が大きいほど，モデルの未知データに対する汎化能力が高いことを意味しており，トピックモデルなどの統計的言語モデルの評価尺度としてよく用いられる．

本実験では，画像に対応付けられた 80%のタグをランダムに抽出してテストセットとして使用し，残りのタグとすべての Visual Word を訓練セットとしてモデルの推定に用いる．テストセットに Visual Word を用いないのは，テストセット対数尤度の値が dense sampling の設定に影響されると考えられるためである．訓練セットを用いてモデルを推定し，テストセットを用いてそのモデルのテストセット対数尤度を測定することで，モデルの精度を評価する．

実験の結果を Fig 8 に示す．h-CorrLDA1 は Visual Word をピボットとした場合の h-CorrLDA を，h-CorrLDA2 はタグをピボットとした場合の h-CorrLDA を表している．h-SymCorrLDA 及び h-CorrLDA のテストセット対数尤度の値が，hLDA のそれに比べて大幅に大きいことがわかる．これは h-SymCorrLDA 及び h-CorrLDA が画像の特徴とタグとの依存性を捉えることができているためだと考えられる．h-CorrLDA2 のテストセット対数尤度の値が h-CorrLDA1 のそれに比べて大きいのは，本実験ではテストセットにタグ情報のみを用いているため，タグをピボットとする h-CorrLDA2 の方が，テストセットに対する汎化能力が高いことを示したと考えられる．また，h-SymCorrLDA のテストセット対数尤度が最も大きい．これは h-CorrLDA がモード間の一方向の依存性しか捉えられないのに対し，h-SymCorrLDA がモード間の相互依存性を捉えられているためだと考えられる．

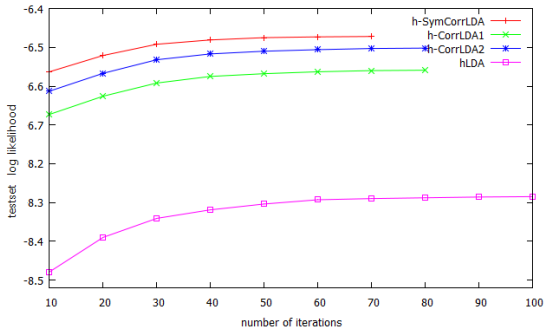

Fig. 8　Testset log likelihood of each hierarchical topic model.

なお，Fig 8 は $C = 1000$ としたときのものである．C の値を動かす実験も行ったが，各モデルのグラフの概観及び性能の優劣については大きな変化は見られなかった．

5.4　正規化相互情報量

正規化相互情報量 (NMI: Normalized Mutual Information) は，正解集合と分類集合の間の依存性を表す量のことで，クラスタリングの評価のために広く用いられている．正解集合 A と分類集合 B との間の NMI は以下の式で表される．

$$\mathrm{NMI}(A,B) = \frac{\mathrm{MI}(A,B)}{\{H(A)+H(B)\}/2}$$

ただし，MI は以下の式で定義される相互情報量である．

$$\mathrm{MI}(A,B) = \sum_{A_i \in A}\sum_{B_j \in B} P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}$$

$P(A_i,B_j)$ は A_i と B_j の同時分布，$P(A_i)$，$P(B_j)$ はそれぞれ A_i と B_j の周辺分布であり，以下の式で表される．

$$P(A_i,B_j) = \frac{|A_i \cap B_j|}{N}$$

$$P(A_i) = \frac{|A_i|}{N},\quad P(B_j) = \frac{|B_j|}{N}$$

ここで，N はデータ (画像) 数である．また，$H(A)$，$H(B)$ はそれぞれ正解集合と分類集合の周辺エントロピーであり，以下の式で表される．

$$H(A) = -\sum_{A_i \in A} P(A_i)\log P(A_i)$$

$$H(B) = -\sum_{B_j \in B} P(B_j)\log P(B_j)$$

NMI は 0 から 1 の値をとり，その値が大きいほど正解集合と分類集合の間に依存関係が現れていることを意味する．

本実験では，あらかじめ与えられている画像クラスに関する階層構造を正解階層とし，各モデルで推定された潜在トピック階層との間の NMI を計算する．ここで，推定された潜在トピック階層において，次の処理を行う．

- 各画像中の単語に割り当たっている各レベルのうち，最も割り当てが多いレベルを選択し，パス中のそのレベルのノードに画像を割り当てる．画像に割り当たっているレベルが同数であれば，ランダムにレベルを選択する．

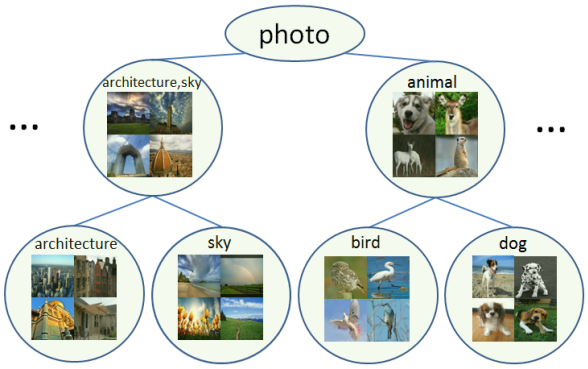

Fig. 9　A subtree generated by h-SymCorrLDA.

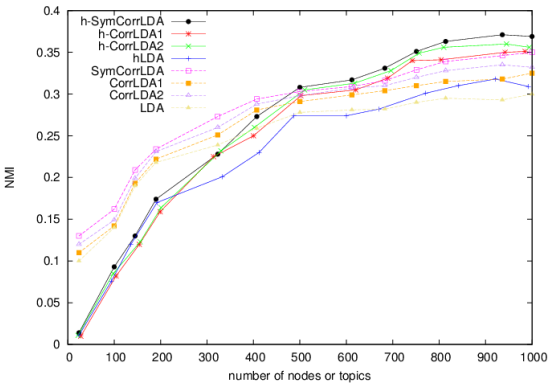

Fig. 10　NMI of each topic model.

この処理を行うことで，ある画像が割り当たるノードを一意に定めることができ，NMI が計算できる．h-SymCorrLDA で推定された画像階層に対し，前述の処理を行った後の木構造の部分木を Fig 9 に示す．例えば右側の部分木では，動物に関するノードの子ノードに鳥に関するノードと犬に関するノードがあり，画像の階層構造が表れていることがわかる．なお，Fig 9 は各ノードに属する画像のうち，例として 4 枚の画像を人手で選んで表示している．

また，本実験では，比較対象として非階層的トピックモデルである LDA，CorrLDA，SymCorrLDA を追加し，トピック数が階層的トピックモデルにおけるノード数に対応すると仮定して比較を行う．ただし，階層的トピックモデルと非階層的トピックモデルはモデルの考え方自体が異なるため，ノード数とトピック数を同一視するのは厳密には適切ではない．

実験の結果を Fig 10 に示す．いずれのモデルについても，ノード数 (トピック数) が 500〜900 付近で曲線の勾配が 0 に近づいている．よって本実験で用いたデータセットにおける適切なノード数 (トピック数) は 500〜900 付近であることがわかる．その適切なノード数 (トピック数) 付近において，h-SymCorrLDA の NMI 値が最も高いことから，提案手法の有効性が示せたと言える．また，ノード数 (トピック数) が大きいとき，階層的トピックモデルの NMI 値が非階層的トピックモデルのそれに比べて大きく，トピックの階層を発見した方がモデルの性能も上がっていることがわかる．これは，非階層的トピックモデルではディリクレハイパーパラメータの影響により，トピックのサンプリング時に全てのトピックが利用される可能性があるが，

階層的トピックモデルでは全ノード (トピック) のうち，木構造中でそのデータに割り当たっているパス中のノードしか利用されない．そのため，階層的トピックモデルではそのデータとはほとんど関係のないトピックが利用されるということが起こりにくく，このことが NMI 値の上昇につながっていると考えられる．

6. おわりに

本論文では，マルチモーダルデータに適用できる新たな階層的トピックモデルとして h-SymCorrLDA を提案し，従来の階層的トピックモデルではできなかったモード間の相互依存性を考慮した潜在トピック階層の推定を実現した．さらに，テストセット対数尤度の測定実験と正規化相互情報量の測定実験を通して提案モデルの有効性を示した．また，トピック-単語多項分布のハイパーパラメータの値の設定方法について，各モードの語彙数を考慮した方法を導入し，その有効性を NMI の測定実験を通して示した．

より詳細な評価実験は今後の課題である．また，画像に対する自動タグ付けなどの評価を行うことが課題として挙げられる．本実験では h-SymCorrLDA の推定にマルチモーダルデータとしてアノテーション付き画像データを用いたが，それに限らず複数のモードを持つデータに対して適用できる．例えば多言語対訳文書への適用が応用として考えられる．また，静止画の系列で表される映像データへの適用も考えられる．

モデル自体に関する今後の課題としては，本実験では推定される潜在トピック階層の高さを指定しているが，自動的にその高さを決定できるようなモデルへと拡張するということが挙げられる．

謝　辞

本研究の一部は，科学研究費補助金基盤研究 (B) (23300039) の援助，および国立情報学研究所の公募型共同研究による．

文　献

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.

[2] Kosuke Fukumasu, Koji Eguchi, and Eric Xing. Symmetric correspondence topic models for multilingual text analysis. *Advances in Neural Information Processing Systems*, Vol. 25, pp. 1295–1303, 2012.

[3] David M Blei, T Griffiths, M Jordan, and J Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, Vol. 16, pp. 106–114, 2004.

[4] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, Vol. 101, No. Suppl 1, pp. 5228–5235, 2004.

[5] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, pp. 5228–5235, 2004.

[6] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134. ACM, 2003.

[7] Li-Jia Li, Chong Wang, Yongwhan Lim, David M Blei, and Li Fei-Fei. Building and using a semantivisual image hierarchy. In *Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3336–3343. IEEE, 2010.

[8] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of 2005 IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 524–531. IEEE, 2005.

[9] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on statistical learning in computer vision*, Vol. 1, pp. 1–2, 2004.

[10] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, Vol. 60, No. 2, pp. 91–110, 2004.