局所特徴のバースト性を考慮した異種メディアの統合的なモデル化

西原　聖志† 　江口　浩二†

† 神戸大学大学院システム情報学研究科情報科学専攻
E-mail: †103x122x@stu.kobe-u.ac.jp, ††eguchi@port.kobe-u.ac.jp

あらまし　近年，社会の高度な情報化にともない，テキストだけでなく画像や映像などの情報が爆発的に増加している．しかし，この大量の情報の中から目的物を探すのは非常に困難であり，データの自動インデクシングなど様々な研究が行われている．なかでも，LDA（Latent Dirichlet Allocation）を代表とするトピックモデルが近年注目されており，その有用性の高さが示されている．例えば，画像データに対するトピックモデルとして CorrLDA があるが，このモデルは単語のバースト性を考慮していないため，映像データに適用した場合その構造をうまくモデル化できないといった問題がある．本論文では，局所特徴のバースト性を考慮した，映像データに対するトピックモデル Corr-DCMLDA を提案する．

キーワード　トピックモデル，潜在的ディリクレ配分法，ギブスサンプリング

Kiyoshi NISHIHARA† and Koji EGUCHI†

† Graduate School of Systems Informatics, Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
E-mail: †103x122x@stu.kobe-u.ac.jp, ††eguchi@port.kobe-u.ac.jp

1. はじめに

インターネットの普及，そして近年におけるソーシャルメディアの急速な発展により，世界中のデータ量が爆発的に増加している．その中にはテキスト情報だけでなく，画像や音声，映像といったデータも数多く存在する．特に，オンラインビデオ共有サービスの代表である Youtube には，一分あたり 24 時間，一日に 60,000 ものビデオがアップロードされており，オンラインビデオは 2013 年のインターネットトラフィックの 91% を占めると推定されている．しかしながら，データの大規模化にともない，目的物を探すのが困難になってきているのも事実であり，現在さまざまなアクセス手段が求められている．このような大規模データを手動でインデクシングするのは不可能であるため，機械学習によるビデオ解析が盛んに研究されている．

機械学習のなかでも最近発展してきているのが，潜在的ディリクレ配分法（LDA：Latent Dirichlet Allocation）をはじめとするトピックモデルと呼ばれる手法であり，その有用性の高さが数多く報告されている．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（トピック）の混合分布から生成される」という考えに基づいた文書モデルで，与えられた文書に潜在するトピックを推定するのに用いられる．またトピックモデルは拡張性が高く，画像データ [1] [2] やネットワークデータ [3] [4] などさまざまなデータにも利用できることが知られている．画像データに対するトピックモデルとして，Correspondence LDA（CorrLDA）[5] がある．このモデルでは画像をベクトル量子化することにより画像をテキストのように扱い，トピックモデルを画像に適用させている．しかし CorrLDA では画像データを想定しているため，映像データに適用した場合，映像の構造をうまくモデル化できないという問題点がある．本論文では，特に局所特徴のバースト性に注目する．ここでいうバースト性とは，元々は文書データの特定の潜在トピックについて，ある文書に一度使われた単語は再びその文書において使われやすいという性質である．これは映像においても同様であり，例えば，ある映像を考えた際，ある潜在トピックに関する画像特徴量は，同一映像において互いに類似し，映像をまたがって類似するとは限らない．バースト性を考慮しない場合，同じトピックではあるが異なる観点で表現されているものを別々のトピックとして扱ってしまい，トピックに対する表現の多様性を捉えられないという問題がある．映像ごとに異なるトピック多項分布を仮定することでバースト性を捉えることができ，トピックに対する表現の多様性を許可することができる．このような考えのもと，本論文では局所特徴のバースト性を考慮した，映像データに対するトピックモデル Correspondence Dirichlet Compound Multinomial LDA（Corr-DCMLDA）を提案する．そしてテストセット対数尤度の測定実験によりモデル評価を行い，提案モデルの有効性を示した．

2. 関連研究

2.1 LDA

LDA（Latent Dirichlet Allocation）[5] はトピックモデルの代表的なものの一つで，文書はある特徴を持った単語の多項分布（トピック）の混合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．LDA ではトピックを推定する際に，文書トピック混合分布とトピック単語多項分布のそれぞれにディリクレ事前分布を仮定する．図.1 にグラフィカルモデルを示し，以下には LDA の文書生成過程を書き下す．ただし，α と β はディリクレ事前分布のハイパーパラメータであり，K と D，N_d はそれぞれトピック数と文書数，文書 d における総単語数を意味する．

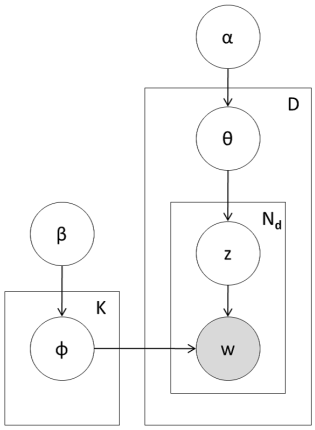

図 1　A graphical model of LDA

（1） ディリクレ事前分布 $Dir(\alpha)$ を用いて，各文書 d に対して多項分布パラメータ θ_d を選択する．

（2） ディリクレ事前分布 $Dir(\beta)$ を用いて，各トピック z に対して多項分布パラメータ ϕ_z を選択する．

（3） 文書 d の単語 $w_{d,i}$ に対して，

- 多項分布 $Mult(\theta_d)$ からトピック $z_{d,i}$ を選択する．
- 多項分布 $Mult(\phi_{z_{d,i}})$ から単語 $w_{d,i}$ を選択する．

ディリクレ事前分布のパラメータ値は大きな値をとるにしたがい，得られる多項分布パラメータの各値が等しくなる．この生成過程は単語のバースト性を考慮していない．例えば，文書に “sports”トピックから得られた単語 “tennis”が含まれているとき，その文書には他にも同じ “sports”トピックからの単語がいくつか含まれていそうである．すなわち，その文書には単語 “tennis”がもう一つ含まれている可能性が高くなる．しかしそれだけではなく，これは “sports”トピックに割り当てられている他のすべての単語に対しても同様である．つまり，単語 “tennis”が存在することにより，間接的にその文書に関係のない単語（たとえば “soccer”，“rugby”など）も出現しやすくなっており，これは望ましくない現象である．後に示す DCMLDA はこのバースト性を考慮したモデルである．

トピックモデルの推定手法としては，変分ベイズ法とギブスサンプリング法 [6] [7] などが知られている．また，ギブスサンプリング法は実装が比較的容易で，コーパスが巨大な際に他の提案手法よりも効率的であることがわかっている．ここではギブスサンプリング法を取り上げ，そのモデル推定の際の式を示す．以降，C^{word} はトピック中の単語のカウント行列を，C^{doc} は文書中のトピックカウント行列を指す．

$$p(z_{d,i}=k|Z_{-(d,i)}, w_{d,i}=v, W_{-(d,i)}, \alpha, \beta) \propto (C^{doc}_{d,k}+\alpha_k)\frac{C^{word}_{v,k}+\beta_v}{\sum_v C^{word}_{v,k}+\sum_v \beta_v} \quad (1)$$

すべての文書における各々の単語に対して，各トピックが割り当てられる確率を上式にしたがって更新する．ランダムな分布から開始して，上の手順を十分な回数だけ繰り返すことで，確率変数間の依存関係を反映した分布を近似的に求めることができる．

2.2 CorrLDA

LDA はテキストデータを対象としたモデルであるが，それを一般物体認識の分野で利用できるように拡張したものが CorrLDA（Correspondence LDA）である．これは画像とテキストデータが対になったデータを対象としたモデルであり，画像分類や自動画像アノテーションなどのコンピュータビジョンの問題にも適用できることが知られている．本来テキストデータに対するモデルであるトピックモデルを画像に適応させるために，テキスト領域においてよく知られている “bag of words” 表現と画像領域における “bag of visual-words” との類似性を導くというアプローチが一般的である．

“bag of visual-words” 表現にするために，SIFT アルゴリズムなどを用いて画像の局所特徴を抽出し，画像のベクトル量子化が行われる．したがって，トピックモデルは相関性のある画像部分の潜在的なグループを発見するために用いられる．また，テキストに関しては画像のトピック割り当て状況も考慮してトピック割り当てを推定する．図.2 にグラフィカルモデルを示し，以下には CorrLDA の文書生成過程を書き下す．ただし，β と β' はそれぞれ画像データ，テキストデータ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り，N_d と N'_d はそれぞれ画像データ，テキストデータにおける総単語数を意味する．

（1） ディリクレ事前分布 $Dir(\alpha)$ を用いて，各画像 d に対して多項分布パラメータ θ_d を選択する．

（2） ディリクレ事前分布 $Dir(\beta), Dir(\beta')$ を用いて，各トピック z に対して多項分布パラメータ ϕ_z, ϕ'_z を選択する．

（3） まず，画像 d 中の visual-word$w_{d,i}$ に対して，

- 多項分布 $Mult(\theta_d)$ からトピック $z_{d,i}$ を選択する．
- 多項分布 $Mult(\phi_{z_{d,i}})$ から単語 $w_{d,i}$ を選択する．

（4） 次に，画像 d のテキスト中の単語 $w'_{d,i}$ に対して，

- 画像中で選択されたトピック $Unif(1,...,N_d)$ の中からトピック $z'_{d,i}$ を選択する．
- 多項分布 $Mult(\phi'_{z'_{d,i}})$ から単語 $w'_{d,i}$ を選択する．

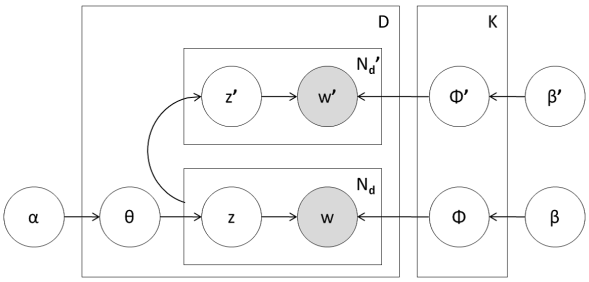

図 2　A graphical model of CorrLDA

次に，CorrLDA のギブスサンプリングの式を以下に示す.

$$p(z_{d,i} = k|Z_{-(d,i)}, w_{d,i} = v, W_{-(d,i)}, \alpha, \beta) \propto (C^{doc}_{d,k,-i} + \alpha_k)\frac{C^{word}_{v,k,-i} + \beta_v}{\sum_v C^{word}_{v,k,-i} + \sum_v \beta_v} \quad (2)$$

$$p(z'_{d,i} = k|Z'_{-(d,i)}, w'_{d,i} = v', W'_{-(d,i)}, \beta') \propto \frac{C^{doc}_{d,k,-i}}{\sum_k C^{doc}_{d,k}}\frac{C^{word}_{v',k,-i} + \beta'_{v'}}{\sum_{v'} C^{word}_{v',k,-i} + \sum_{v'} \beta'_{v'}} \quad (3)$$

なお，$-i$ という表記は v_i がデータから除去されたということを意味している.

2.3　DCMLDA

従来のトピックモデルではバーストに現れる単語の傾向（文書または本に一度使われた単語は再びその文書または本で使われやすいという言語の基本特性）を捉えられないという欠点がある．バースト性を文書レベルで捉える研究 [8] と本レベルで捉える研究 [9] があるが，ここでは後者を取り上げる．DCM 分布（Dirichlet Compound Multinomial）を用いてこのバースト現象をモデル化したものが DCMLDA（Dirichlet Compound Multinomial LDA）である．DCMLDA モデルではバースト性を考慮するために本ごとに異なるトピック多項単語分布を仮定する．また，本ごとにトピックモデルを計算するため，LDA に比べて計算量が少なく，非常に巨大なコレクションにも適用できるのが特徴である.

DCM 分布の本に対する有効性は Mimno らによって報告されている．例えば，ある本を考えた際，その本のページは無作為に選ばれた他の本のページよりも同じ本の他のページにトピック的にずっと類似しているという性質がある．DCMLDA では本ごとにトピックが独立しているため,「本質的に同じトピックについて書かれているが，少し違った観点で論じられている」といった，本ごとのバースト性をうまくモデル化することができる．図.3 に Mimno らによる DCMLDA のグラフィカルモデルを示し，以下にその生成過程を書き下す.

（1）　各本 b において，ディリクレ事前分布 $Dir(\alpha_b)$ を用いて，各ページ d に対して多項分布パラメータ θ_{bd} を選択する.

（2）　各トピック z のディリクレ事前分布 $Dir(\beta_z)$ を用いて，本 b の各トピック z に対して多項分布パラメータ ϕ_{bz} を選択する.

（3）　本 b，ページ d の単語 $w_{d,i}$ に対して，

- 多項分布 $Mult(\theta_{bd})$ からトピック $z_{d,i}$ を選択する.
- 多項分布 $Mult(\phi_{bz_{d,i}})$ から単語 $w_{d,i}$ を選択する.

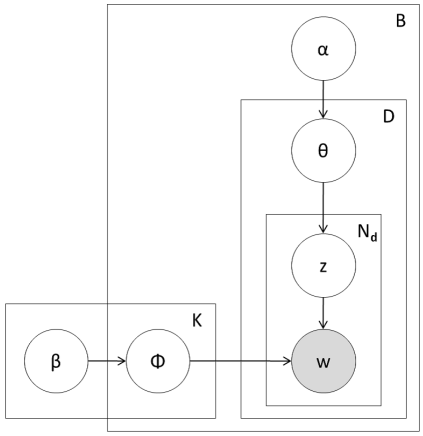

図 3　A graphical model of DCMLDA

LDA では，各トピック k に対して，一つの多項分布 ϕ_k がディリクレ事前分布 ($Dir(\beta)$) から導出され，すべての文書に利用される．DCMLDA では，各トピック k と各本 b に対して新しい多項単語分布 ϕ_{bk} を仮定する．さらに，トピック k ごとに非対称ディリクレ事前分布 $Dir(\beta_k)$ を仮定する．各本ごとのトピック分布をもつことにより異なる本内の同じトピック中の各単語の確率変動（すなわちバースト現象）を捉えることができる.

2.4　映像データに対するトピックモデル

映像データに対するトピックモデルもいくつか提案されている.

映像データにはたくさんの情報が含まれており，例えば，音声データやカラーヒストグラム，局所画像領域のテクスチャなどがある．映像データに含まれている情報をできる限り保存するために，これらの表現は一般的に高次元であるが，これは非効率であり大量のメモリも必要とする．Wanke らはトピックモデルを用いて映像データの意味圧縮に取り組んでおり，結果として，予測精度を著しく失うことなく従来のヒストグラム表現に比べて 20 : 1，他の次元縮小法と比べて少なくとも 2 : 1 の圧縮比を達成している [10]．Wanke らの研究では,「映像中ではトピックの変化はショットの切替と同時に起こらなければならない」という映像の時間的構造を利用しており，映像データに対してトピックモデルと隠れマルコフモデルの組み合わせを適用している．また，トピックモデルには PLSA を用いており，“bag of visual words”表現を使って画像に適応させている．一方，Souvannavong らは映像データにおける物体検索やシーン分類のようなタスクに向けた LSA を開発している [11] [12] [13]. これらのどの研究においても，画像の特徴量をベクトル量子化し，マルチモーダル情報をうまく組み合わせてトピックモデルを適用している．しかしながら，これらは映像におけるバースト性を考慮していないため,「同じトピックについて異なる観点で表現する」といったトピックに対する多様性を捉えられない.

3. 提案モデル

3.1 Corr-DCMLDA

本論文では，局所特徴のバースト性を考慮した，映像データに対するトピックモデル Corr-DCMLDA（Correspondence Dirichlet Compound Multinomial LDA）を提案する．映像データは画像データと音声データの集合であり，データとしては映像から抽出したキーフレーム画像，音声を文字に書き起こした音声書き起しデータが対象である．キーフレーム画像を visual-word 特徴で表現し，それに対応する音声書き起こしテキストとともに潜在トピックを推定することで，統合的なモデル化を試みる．visual-word とは画像データにおける離散特徴量のことであり，テキストデータにおける単語のような性質を持っている．

本論文では特に，同種のオブジェクトであるにも関わらず，映像によって異なる visual-word 特徴が出現するバースト現象を考慮する．図.4 に Corr-DCMLDA のグラフィカルモデルを示し，以下にその生成過程を書き下す．

（1） 各映像 b において，ディリクレ事前分布 $Dir(\alpha_b)$ を用いて，各キーフレーム d に対して多項分布パラメータ θ_{bd} を選択する．

（2） 各トピック z のディリクレ事前分布 $Dir(\beta_z), Dir(\beta'_z)$ を用いて，各映像 b の各トピック z に対して多項分布パラメータ ϕ_{bz}, ϕ'_{bz} を選択する．

（3） まず，映像 b，キーフレーム d 中の visual-word$w_{d,i}$ に対して，

- 多項分布 $Mult(\theta_{bd})$ からトピック $z_{d,i}$ を選択する．
- 多項分布 $Mult(\phi_{bz_{d,i}})$ から単語 $w_{d,i}$ を選択する．

（4） 次に，映像 b，キーフレーム d の音声データ中の単語 $w'_{d,i}$ に対して，

- 画像中で選択されたトピック $Unif(1, ..., N_d)$ の中からトピック $z'_{d,i}$ を選択する．
- 多項分布 $Mult(\phi'_{bz'_{d,i}})$ から単語 $w'_{d,i}$ を選択する．

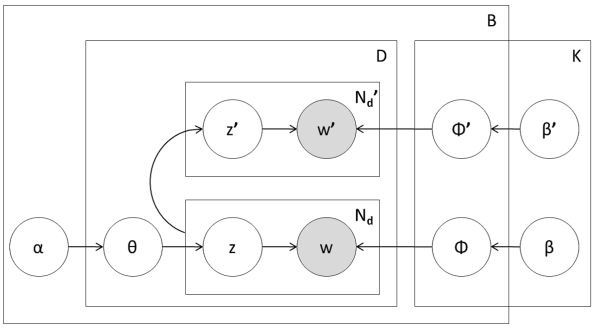

図 4 A graphical model of Corr-DCMLDA

本来，画像データとテキストデータが対になったデータに対しては CorrLDA が有効であるが，これは映像データには当てはまらない．CorrLDA の重要な問題点は，LDA と同じくバースト性を考慮していないためにコーパスの構造を大きく無視してしまっているところである．つまり，CorrLDA を映像データに適用した場合，映像の構造を無視してしまうのである．そこで，DCM 分布を用いてその問題を解決する．バースト性は画像の局所特徴に対しても同様であり，2.3 節で述べた「本とページ」が「映像と画像」に対応するため，DCM 分布が利用できる．各映像は自身のトピックを持ち，そのトピックは映像間で共有されたものではない．そのため本モデルでは「本質的に同じトピックについての映像だが，映像によって違った観点で表現されている」といった，映像データにおけるバースト性をうまくモデル化することができる．つまり，各映像がはっきりと異なった観点を持ちながら映像の類似性を発見できるのである．また，映像コーパス全体でモデルを推定する CorrLDA とは異なり，Corr-DCMLDA では映像ごとに独立にモデル推定を行うためモデル推定にかかる時間を抑えることができる．

Corr-DCMLDA のモデル推定のための全体的な流れを Fig.5 に示す．詳細については次節以降で述べる．

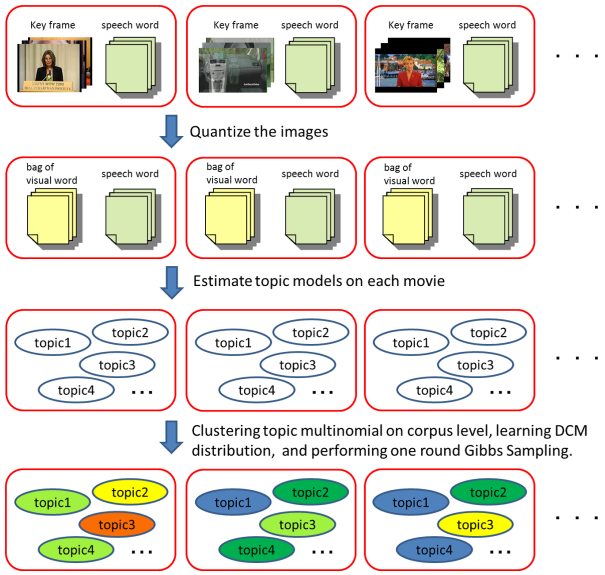

図 5 A flow of estimating Corr-DCMLDA

3.2 前 処 理

映像データは画像と音声の集合であるため，モデル推定にあたりなんらかの前処理が必要である．まず，動画からキーフレームを抽出して画像を取り出し，音声データを音声認識を用いて書き起し，テキストデータ (speech-word) とする．次に，画像にトピックモデルを適用するために画像のベクトル量子化を行う必要がある．これには SIFT アルゴリズムを用いる．SIFT アルゴリズムには自動特徴点抽出と SIFT 記述子抽出の 2 つの方法があり，前者は自動的に画像の特徴的な点を抽出するものであり，後者は指定された点に対して SIFT 記述子を抽出するものである．SIFT で得られる特徴量は回転やスケールにロバストであり，画像認識の分野で非常に注目されている技術である．本論文では自動特徴点抽出ではなくリージョンを指定して特徴量を抽出する．自動特徴点抽出の場合，特徴点数が極めて少ない画像が出てきたり，各画像で特徴点数に大きなばらつきが出てしまい，モデル推定の際に大きな障害となってしまうためである．まず各画像を等しい解像度に揃えた上で，

リージョン指定により各画像に対して 10×10 ピクセル間隔で SIFT 記述子を抽出し，スケールは $10 \sim 30$ でランダムにサンプリングする．これによりどの画像においても十分な特徴量を得ることができる．なお，ピクセル間隔およびスケールは Li らの研究を参考にしている [2]．得られた局所特徴を k-means アルゴリズムを用いてクラスタリングを行い visual-word とすることで，画像のベクトル量子化は完了する．なお，経験的に visual-word 数は 1000 とした．

3.3 トピックのクラスタリング

各映像に対して独立にトピックモデルを計算した後，そこで得られたトピックをコーパスレベルのトピックにクラスタリングし，それらのクラスタから DCM 分布を学習する．また，各映像に対してトピックモデルを計算する際，Minka [14] の不動点反復法を用いてハイパーパラメータ α を推定する．

$$\alpha_k^{new} = \alpha_k \frac{\sum_d \Psi(C_{d,k}^{doc} + \alpha_k) - \Psi(\alpha_k)}{\sum_d \Psi(\sum_k C_{d,k}^{doc} + \sum_k \alpha_k) - \Psi(\sum_k \alpha_k)} \quad (4)$$

DCM 分布を学習するためにコーパスレベルでトピック単語多項分布のクラスタリングを行う．これは k-means アルゴリズムを用いる．また，クラスタリングのための距離尺度には Jensen-Shanonn（JS）ダイバージェンスを用いる．JS ダイバージェンスは，二つの分布 p，q に対して平均分布 $r = (p+q)/2$ を定義すると，p と r の Kullback-leibler（KL）ダイバージェンスと q と r の KL ダイバージェンスの平均である．JS ダイバージェンスを用いる理由として 2 点挙げられる．まず一つ目に，JS ダイバージェンスには対称性があり，比較される順序は重要ではないということである．二つ目に，もし p にあって q にない単語 w があったとき，$p(w)\log q(w)$ 項はゼロの対数も含むため，このとき KL ダイバージェンスは不定になるのに対して JS ダイバージェンスはそうではないことである．これは重要な留意事項である．両方のトピックに出現する単語の確率のみを考慮することで JS ダイバージェンスを以下のように変形することができる．

$$1 + \frac{1}{2} \sum_{w \in \{p \cup q\}} p(w)\log q(w) + q(w)\log q(w) - (p(w)+q(w))\log\frac{p(w)+q(w)}{2} - p(w) - q(w) \quad (5)$$

またトピックのクラスタリングにおいては，画像データのトピック単語多項分布 p_{im}，および音声書き起こしテキストのトピック単語多項分布 q_{sp} を線形結合し，一つの分布 r_{mm} として扱う．このとき，λ を用いて次式のような重み付けを行う．

$$r_{mm} = \lambda p_{im} + (1-\lambda) q_{sp}$$

線形結合された多項分布 r でトピックのクラスタリングを行う．本論文では，$\lambda = \{0, 0.25, 0.5, 0.75, 1\}$ の 5 通りで実験を行った．

トピックのクラスタを特定した後，各クラスタに対して DCM 分布を学習するためにそのクラスタに割り当てられているトピックを使う．これにも α の推定と同様，Minka の不動点反復法を用いる．

$$\beta_{kv}^{new} = \beta_{kv} \frac{\sum_k \Psi(C_{v,k}^{word} + \beta_{kv}) - \Psi(\beta_{kv})}{\sum_k \Psi(\sum_v C_{v,k}^{word} + \sum_v \beta_{kv}) - \Psi(\sum_v \beta_{kv})} \quad (6)$$

今回は各クラスタに対して 10 回繰り返すことで DCM パラメータ β_k を学習した．最後に，学習した DCM パラメータを各映像に伝播させるために，各映像に対してギブスサンプリングを 1 ラウンド行うことで，これを実行する．このとき，DCM パラメータと，各映像で推定したハイパーパラメータ α を用いる．

4. 実　　験

提案手法の有効性を示すために，モデル推定実験を行った．

4.1 データセット

今回の実験では，MediaEval で提供されたデータを用いる．具体的には，映像数 247，1726 の二つのデータセットを用いる（以降，それぞれをデータセット 1，データセット 2 とする）．データセットの詳細を表.1 に示す．なお，visual-word 数は経験的にどちらも 1000 に設定している．

表 1　Summary of datasets used

	dataset1	dataset2
the number of movies	247	1726
the number of images	9330	59624
the number of speech-words	6291	17595

4.2 Minka の不動点反復法

トピックをクラスタリングした後，DCM パラメータを学習するために 3.3 節で述べた通り，Minka の不動点反復法を用いて，各クラスタに対して 10 回繰り返す．トピック数 $T = 5$，クラスタ数 $C = 5$ のときの，各繰り返しごとの DCM パラメータに関する残差二乗和のマイクロ平均を図.6，図.7 に示す．クラスタ i のメンバ数を n_i，残差二乗和を x_i とすると，マイクロ平均 $micro$ は次式で計算できる．

$$micro = \frac{\sum_{i=1}^{C} n_i x_i}{\sum_{i=1}^{C} n_i} \quad (7)$$

図.6，図.7 より，visual-word，speech-word ともに 10 回の繰り返しで収束していることがわかる．トピック数，クラスタ数が変化してもこの傾向は同様であった．

4.3 テストセット対数尤度の測定実験

4.3.1 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テストセット対数尤度を用いる．テストセット対数尤度が大きければ大きいほど，その推定モデルの精度が高いことを意味する．本実験では，コーパス全体にわたり無作為に選択した speech-word の 90% を開発セットとし，残りの 10% をテストセットとした．なお，visual-word は本来画像データの連続的な特徴量を離散的な値として抽出したものであり，SIFT での点の選び方で変化してしまうなど，テ

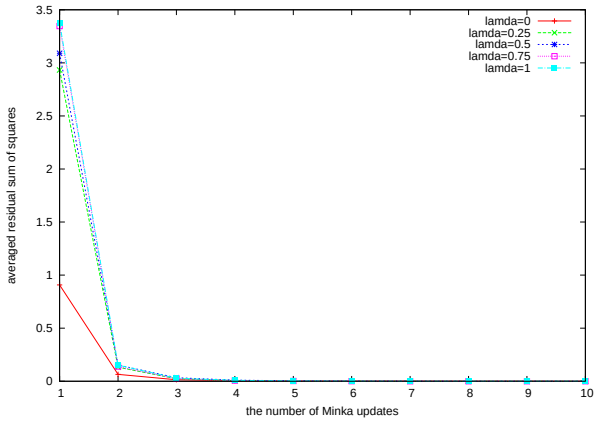

図 6　Micro average of residual sum of squares on DCM parameters in the case of visual words

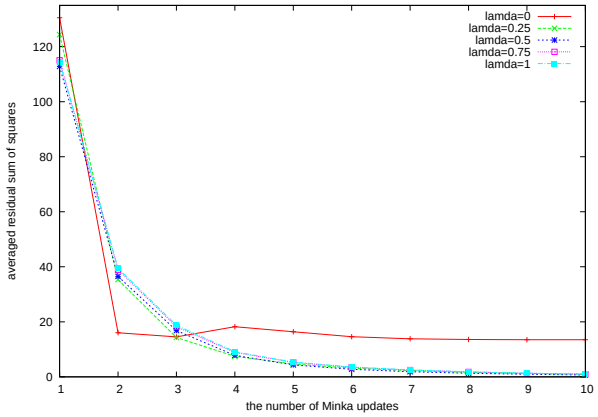

図 7　Micro average of residual sum of squares in the case of speech words

ストセットとしては不適であるため，speech-word のみのテストセット対数尤度でモデル評価を行う．なお speech-word において，あらかじめ 418 語のストップワードを除去している [15]．さらに全コーパスにわたる出現回数が 5 回未満の単語も除去している．また，テストセット対数尤度の測定において，90% の開発セットに現れず，10% のテストセットにのみ現れる単語も除去している．

Corr-DCMLDA における，テストセット x^{test} の尤度は次式で求めることができる．

$$p(x^{test}) = \prod_{ij} \sum_{k} \frac{C'^{doc}_{d,k}}{\sum_k C'^{doc}_{d,k}} \frac{C^{word}_{v,k} + \beta^{DCM}_{kv}}{\sum_v C^{word}_{v,k} + \sum_v \beta^{DCM}_{kv}}$$

ここで，C'^{doc} は画像中の visual-word のトピックカウント行列を指し，${\beta^{DCM}}_{kv}$ はトピックのクラスタリングによって学習した DCM 分布のハイパーパラメータである．

4.3.2　比 較 対 象

本実験では，二つの比較対象を用意した．一つ目は，映像という構造を考慮せず，コーパス全体にわたって適用した CorrLDA である．CorrLDA は画像データとテキストデータが対になったデータを対象としたトピックモデルであるためそのまま適用できる．二つ目は DCMLDA である．visual-word と speech-word を同じ種類の単語とみなし，映像に対して DCMLDA モデルを適用する．これら二つのモデルと提案モデルをテストセット対数尤度で比較することで提案モデルの有効性を示す．

4.3.3　実 験 設 定

本実験では，トピック数 $T = 5, 10, 20, 40, 80$ として実験を行った．提案モデル，および比較対象のモデルはすべてギブスサンプリング法により推定した．ハイパーパラメータ α は全てのモデルで 50 ギブススイープ終了毎に，Minka の不動点反復法を 5 回繰り返すことで推定した．ハイパーパラメータ β は推定を行わず，$\beta = 0.1$ に固定した．

また，DCM 分布の学習におけるトピックのクラスタリングにおいて，クラスタ数 C をトピック数 T に揃えた場合と，クラスタ数をトピック数とは異なる $C = 5, 10, 20, 40, 80$ の場合で実験を行った．

4.3.4　実 験 結 果

実験の結果を表と図にまとめた．CorrLDA のデータセット 1 での結果を表.2，図.8 に，データセット 2 での結果を表.3 に示す．DCMLDA の結果を表.4，図.9 に示す．また，表.5，表.6 はそれぞれ，Corr-DCMLDA をトピック数とクラスタ数が等しい場合，異なる場合のデータセット 1 での実験結果であり，それをグラフ化したものが図.10，図.11 である．Corr-DCMLDA のデータセット 2 での結果は表.7，図.12 である．

全ての結果からわかる通り，Corr-DCMLDA が二つの比較対象よりも良い結果を示した．また，表.5 より，トピック数とクラスタ数が同じ場合，$T = (C =)5$ のときが最も良い値であった．表.2 におけるトピック数 $T = 1235$ はトピック数 $T = 5$ とデータセット 1 の映像数 247 を乗じたものである．これは，Corr-LDA と Corr-DCMLDA で扱うトピック数を実質的に等しくし，映像コーパス全体でモデル推定を行った場合と映像ごとにモデル推定を行った場合とを比較するための実験である．結果を見ると，映像全体にわたってモデルを推定する CorrLDA よりも映像のバースト性を考慮した Corr-DCMLDA のほうがテストセット対数尤度の値が高いことがわかる．これはトピック数，クラスタ数のどちらかが影響したと考えられるため，それを検証したのが表.6 である．表.6 において，どのトピック数においても，クラスタ数が $C = 5$ のときに最も良い値を示している．この結果より，トピックのクラスタリングにおいておおまかなクラスタリングのほうが有用であると考えられる．

Corr-DCMLDA の λ に関してはどの実験においても $\lambda = 0$，つまりトピックのクラスタリングにおいて speech-word のみでクラスタリングを行なったときに最も良い結果になった．これはテストセットを speech-word のみとしているためであり，speech-word のみを考慮したクラスタリングのほうがテストセット対数尤度が高くなるのは自明である．検索や分類問題における性能に対する λ のふるまいを調査することは今後の課題である．なお，λ のどの場合においても Corr-DCMLDA が他の比較対象より良い結果を与えていることは明らかである．

表 2　CorrLDA with dataset1

number of topics	test-set log-likelihood
5	-7.7619
10	-7.7520
20	-7.7178
40	-7.6706
80	-7.6258
1235	-7.7768

表 3　CorrLDA with dataset2

number of topics	test-set log-likelihood
40	-8.1385
80	-8.1169
160	-8.1112

表 4　DCMLDA with dataset1

number of topics	test-set log-likelihood
5	-9.8775
10	-9.5620
20	-9.3318
40	-9.1371
80	-9.0041

表 5　Corr-DCMLDA with dataset1 when the number of topics is the same as the number of clusters

number of topics	number of clusters	λ				
		0	0.25	0.5	0.75	1
5	5	-7.1795	-7.2280	-7.2325	-7.2304	-7.2307
10	10	-7.2013	-7.3008	-7.3053	-7.3071	-7.3063
20	20	-7.2229	-7.3826	-7.3876	-7.3882	-7.3852
40	40	-7.2481	-7.4584	-7.4598	-7.4632	-7.4693
80	80	-7.2890	-7.5451	-7.5415	-7.5554	-7.5508

表 6　Corr-DCMLDA with dataset1 when the number of topics is not the same as the number of clusters

number of topics	number of clusters	λ				
		0	0.25	0.5	0.75	1
5	5	-7.1795	-7.2280	-7.2325	-7.2304	-7.2307
	10	-7.2060	-7.2652	-7.2788	-7.2741	-7.2776
	20	-7.2154	-7.3191	-7.3202	-7.3289	-7.3252
10	5	-7.1984	-7.2490	-7.2623	-7.2515	-7.2626
	10	-7.2013	-7.3008	-7.3053	-7.3071	-7.3063
	20	-7.2177	-7.3597	-7.3644	-7.3566	-7.3630
20	5	-7.1977	-7.2799	-7.2764	-7.2773	-7.2773
	10	-7.2058	-7.3186	-7.3268	-7.3272	-7.3282
	20	-7.2229	-7.3826	-7.3876	-7.3882	-7.3852

表 7　Corr-DCMLDA with dataset2 when the number of topics is not the same as the number of clusters

number of topics	number of clusters	λ				
		0	0.25	0.5	0.75	1
5	5	-7.5880	-7.6713	-7.6839	-7.6536	-7.6648
	10	-7.6069	-7.7345	-7.7565	-7.7463	-7.7562
	20	-7.6130	-7.7888	-7.8018	-7.8000	-7.8031
10	5	-7.6051	-7.6894	-7.6825	-7.7034	-7.7122
	10	-7.6112	-7.7859	-7.7907	-7.7826	-7.7839
	20	-7.6277	-7.8334	-7.8415	-7.8395	-7.8410
20	5	-7.6138	-7.6997	-7.7167	-7.7052	-7.7179
	10	-7.6227	-7.7920	-7.7937	-7.7922	-7.7915
	20	-7.6312	-7.84430	-7.8397	-7.8497	-7.8482

5. おわりに

本論文では，局所特徴のバースト性を考慮した，映像データに対するトピックモデル Corr-DCMLDA を提案した．映像も文書と同様にバースト性を持ち,「ある潜在トピックに関する画像特徴量は同一映像において互いに類似し，映像をまたがって類似するとは限らない」と考えることができる．提案モデルでは，トピックに対する表現の多様性を許可するために，映像ごとに異なるトピック多項分布を仮定し，コーパス全体でトピックのクラスタリングを行うことでこれを実現している．提案モデルの有効性に関して，テストセット対数尤度による測定実験

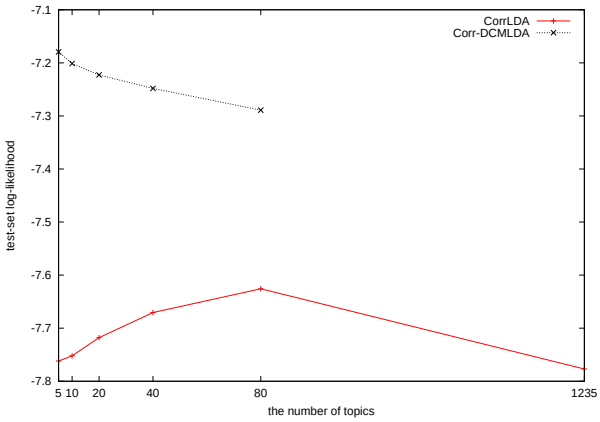

図 8　test-set log-likelihood of CorrLDA with dataset1

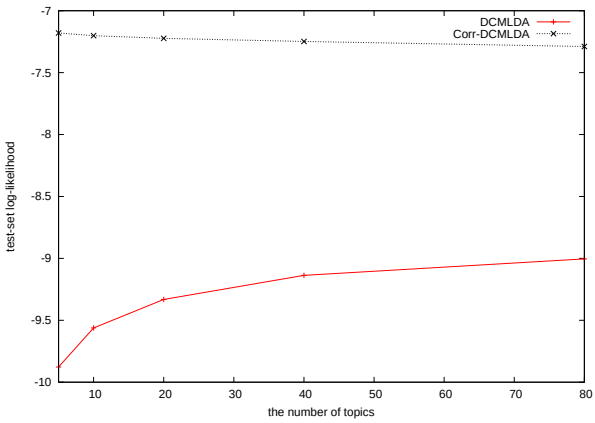

図 9　test-set log-likelihood of DCMLDA with dataset1

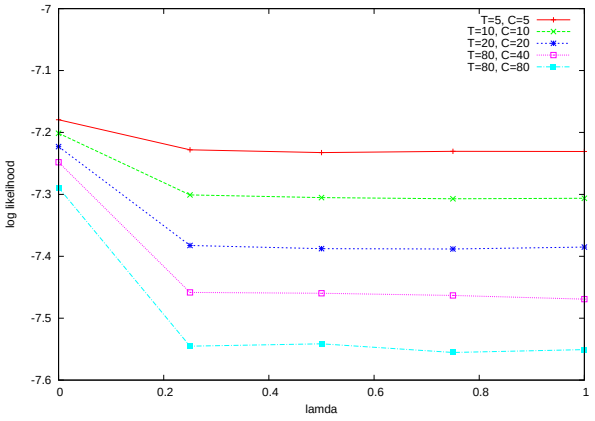

図 10　test-set log-likelihood of Corr-DCMLDA with dataset1 when the number of topics is the same as the number of clusters

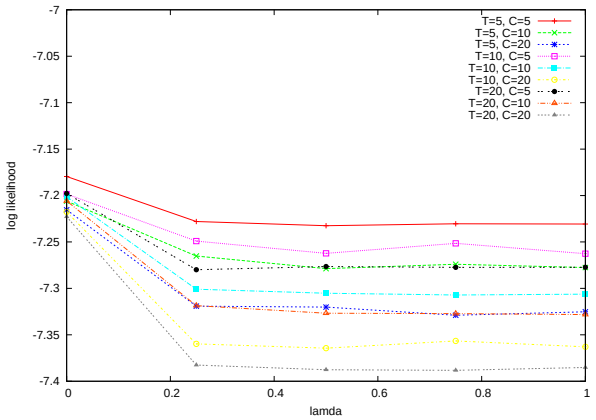

図 11 test-set log-likelihood of Corr-DCMLDA with dataset1 when the number of topics is not the same as the number of clusters

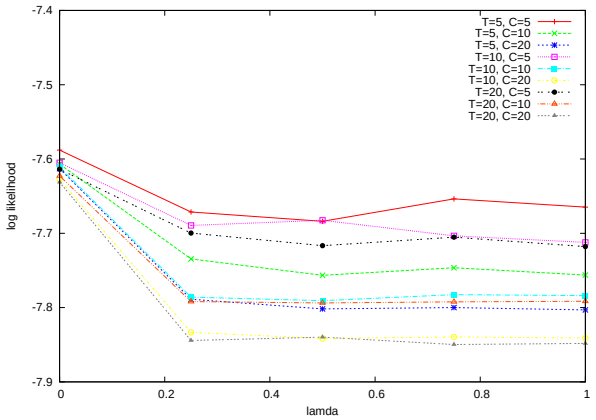

図 12 test-set log-likelihood of Corr-DCMLDA with dataset2 when the number of topics is not the same as the number of clusters

を行い，二つの既存手法と比較した．コーパス全体でのトピックのクラスタリングにおいてはおおまかなクラスタリングのほうが結果が良好であり，これはトピックの表現の多様性を捉える上で重要な側面である．文書または本のバースト性を考慮することでより良いトピックモデルを推定できることは知られていたが，映像データにおいても，局所特徴のバースト性を考慮することでうまくモデル化できることがわかった．

本論文ではテストセット対数尤度によるモデル評価のみ行なったため，ジャンル分類，クラス分類，映像検索などのタスクの性能評価は今後の課題である．また，タスク評価における λ の評価も課題として挙げられる．なお，提案モデルでは画像データとテキストデータを含む映像データを対象としているため，音声データだけでなくソーシャルタグなどのテキストデータの場合の拡張モデルも考えられる．

謝　　辞

本研究の一部は，科学研究費補助金基盤研究 (B)（20300038, 23300039）の援助による．

文　　献

[1] Blei, D. M. and Jordan, M. I.: Modeling annotated data, in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pp. 127–134, ACM (2003).

[2] Li, F.-F. and Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories, in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pp. 524–531, IEEE Computer Society (2005).

[3] Zhang, H., Qiu, B., Giles, C. L., Foley, H. C. and Yen, J.: An LDA-based community structure discovery approach for large-scale social networks, *IEEE Intelligence and Security Informatics*, pp. 200–207 (2007).

[4] Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P.: Mixed membership stochastic block models, *Journal of Machine Learning Research*, Vol. 9, pp. 1981–2014 (2008).

[5] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[6] 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).

[7] Steyvers, M. and Griffiths, T.: *Handbook of Latent Semantic Analysis*, chapter 21: Probabilistic Topic Models, Lawrence Erbaum Associates, Mahwah, New Jersey and London (2007).

[8] Doyle, G. and Elkan, C.: Accounting for burstiness in topic models, in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 281–288, ACM (2009).

[9] Mimno, D. and McCallum, A.: Organizing the OCA: learning faceted subjects from a library of digital books, in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '07, pp. 376–385, ACM (2007).

[10] Wanke, J., Ulges, A., Lampert, C. H. and Breuel, T. M.: Topic models for semantics-preserving video compression, in *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pp. 275–284, ACM (2010).

[11] F. Souvannavong, B. M., L. Hohl and Huet, B.: Structurally Enhanced Latent Semantic Analysis for Video Object Retrieval, pp. 859–867, Vision, Image and Signal Processing, IEE Proceedings (2005).

[12] Souvannavong, F., Merialdo, B. and Huet, B.: Latent semantic analysis for an effective region-based video shot retrieval system, in *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, MIR '04, pp. 243–250, ACM (2004).

[13] F. Souvannavong, B. M. and Huet, B.: Latent Semantic Indexing for Semantic Content Detection of Video Shots, Vol. 3, pp. 1783–1786, Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on (2004).

[14] Minka., T.: Estimating a Dirichlet distribution, Vol. 2000, pp. 1–13, Technical report, MIT (2000).

[15] Callan, J. P., Croft, W. B. and Harding, S. M.: The INQUERY Retrieval System, in *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pp. 78–83, Valencia, Spain (1992).