DEIM Forum 2012 E2-2

周辺文を考慮するトピックモデルを用いた評価側面の推定

小西　卓哉† 　木村　文則†† 　前田　亮††

† 立命館大学理工学研究科　〒 525–8577 滋賀県草津市野路東 1–1–1
†† 立命館大学情報理工学部　〒 525–8577 滋賀県草津市野路東 1–1–1
E-mail: †cm005069@ed.ritsumei.ac.jp, ††{fkimura@is, amaeda@media}.ritsumei.ac.jp

あらまし　本稿では，評価文書の評価側面の推定について検討する．評価側面とは，商品やサービスが評価される観点や特徴を表し，例えば工業製品ならば，そのデザインや性能，耐久性を考えることができる．これらを文書中の記述から推定することは，評価文書を活用する様々な場面で有益だと考えられる．本稿では，評価側面をトピックモデルによる推定を検討し，distance dependent Chinese restaurant process を応用した新しいモデルを提案する．また，パープレキシティを用いて定量的な評価で従来手法と比較し，その有効性を検証する．
キーワード　評判分析，自然言語処理，データマイニング

Estimating Ratable Aspects Using the Topic Model Accounting for Neighboring Sentences

Takuya KONISHI†, Fuminori KIMURA††, and Akira MAEDA††

† Graduate School of Science and Engineering, Ritsumeikan University
1–1–1 Noji–Higashi, Kusatsu, Shiga 525–8577, Japan
†† College of Information Science and Engineering, Ritsumeikan University
1–1–1 Noji–Higashi, Kusatsu, Shiga 525–8577, Japan
E-mail: †cm005069@ed.ritsumei.ac.jp, ††{fkimura@is, amaeda@media}.ritsumei.ac.jp

Abstract　We consider ratable aspects of online reviews. Ratable aspects indicate properties of items and services evaluated by reviewers, for instance, reviews about some electronic items are described about design, performance, or durality of each product. In this paper, we estimate these aspects from online user reviews using a topic model. While there is a sophisticated model proposed in previous work, we develop the new model applying the distance dependent Chinese restaurant process. We verify whether our proposed model improves perplexity compared to the previous work in empirical results.
Key words　Sentiment Analysis, Natural Language Processing, Data Mining

1. はじめに

近年，様々な Web 上のデータの解析が進んでいる．本研究では，その中でも商品やサービスの意見が記述された文書である，評価文書に着目する．評価文書は Amazon のようなオンラインショッピングサイトなどを介して，多数のユーザから投稿されており，購入者の生の声を反映した情報源の一つとして活用されている．これらは日々増加を続けており，また様々なサービスから発信されているため，文書を効率的に収集し，そこから必要な情報を抽出することが重要となる．

本研究では，これら評価文書中の記述から評価側面を推定することを検討する．評価側面とは商品やサービスが評価される観点や特徴を表すものである．本稿では，事例としてデジタルカメラの評価文書の例を挙げるが，デジタルカメラならば，その外観や造りの良さ評価する「デザイン」，カメラの画質の良さを表す「画質」，電池の持ちを表す「バッテリー」といった評価側面が考えられる．これら評価側面によって文書中の記述を特徴づけることができれば，評価文書から情報を抽出する様々な場面に応用できる．例えば，評価文書の極性と組み合わせることで，評価側面ごとに文書中の記述の極性を判別することや，評価側面ごとに文書を要約することで，商品の特徴を簡潔に提示するといった応用が考えられる．

本研究では，この評価側面を文書中から推定するために，トピックモデルの利用を検討する．トピックモデルは文書の生成過程を確率的生成モデルによって表すことで，文書集合内の潜在的なトピックを，単語間の関連性として推定する手法である．

このトピックは本研究で検討する評価側面と対応すると考えられ，トピックモデルを適用することで，その推定が期待できる．

評価文書にトピックモデルを適用する際，評価文書の類似性が問題になる．同一カテゴリの評価対象について書かれた評価文書は，それぞれ文書単位で使用される単語の傾向が非常に類似している．通常のトピックモデルは文書単位で Bag-of-words を仮定するため，文書単位のモデル化では意味あるトピックの推定が困難になる．

この問題を解決する最も単純な方法は，文書よりローカルな単位である，文単位で Bag-of-words を仮定することでトピックを推定する方法である．この方法によってある程度意味あるトピックの推定が可能であるが，各文中の単語が少数であるという新たな問題が生まれる．各文のトピックを推定する際，特徴語の乏しい文では，文書中の文脈を無視したトピックが推定されてしまう場合も考えられる．

これまでにトピックモデルを評価文書のモデル化に応用した研究に，Multi-Grain Latent Dirichlet Allocation （MG-LDA）が挙げられる [1]．この手法では，スライディングウィンドウという要素を導入することで，単純な 1 文ごとのモデル化ではなく，周辺の文を考慮したモデル化が可能にしている．本稿では distance dependent Chinese restaurant process（dd-CRP）[2] を応用した，評価側面推定のための新たなモデルを提案する．提案手法と既存手法を比較し，その有効性を検証する．

本稿の構成は以下の通りである．まず 2 章で本研究の問題設定について述べる．3 章で先行研究である MG-LDA を簡略化した，提案手法と比較する simplified MG-LDA を導入した後，先行研究を考察する．4 章で新たに提案するモデルについて述べる．5 章でパープレキシティを用いた定量的評価の結果を示す．最後に 6 章でまとめと今後の課題について述べる．

2. 問題設定

本章では，本研究で手法を適用する評価文書に関して 2 点ほど補足する．まず，同一の商品・サービスカテゴリについて書かれた評価文書の集合をモデル化の対象とする．例えば，家電製品ならば，デジタルカメラ，ノートパソコン，テレビといったカテゴリが考えられる．これらカテゴリごとに評価文書を収集し，トピックモデルを適用することで，各カテゴリの評価側面をトピックとして推定する．

次に，評価文書集合の数学的表現には，文単位で Bag-of-words を仮定する．代表的なトピックモデルである Latent Dirichelt Allocation（LDA）[3] をはじめ，通常のトピックモデルでは，文書単位で Bag-of-words を仮定することが一般的であるが，評価文書集合のモデル化には必ずしも適さない．1 章で述べたように，同一カテゴリに属する評価文書集合は，各々の文書の類似性が高くなることが特徴である．デジタルカメラの評価文書集合ならば，どの文書も同じくデジタルカメラの評価について記述されているため，どの文書にも同じ単語が使用される傾向がある．これにより，文書単位の単語の出現頻度を利用すると，文書間で単語の出現傾向の差が小さくなるため，意味あるトピックの推定が困難となる．本研究では，このような背景から文単位の Bag-of-words を仮定する．ただし，文書中の文の順序については既知とし，モデル化に利用する．先行研究である MG-LDA も同様の仮定をおいている．

Algorithm 1 Generative process of sMG-LDA

for all topic t (=1...T) **do**
  Draw $\boldsymbol{\phi}_t \sim Dir(\boldsymbol{\beta})$
end for
for all document j (=1...D) **do**
  for all sentence s of document j (s=1... S_j) **do**
    Draw $\boldsymbol{\psi}_{js} \sim Dir(\boldsymbol{\gamma})$
  end for
  for all sliding window m of document j (m=1...M_j) **do**
    Draw $\boldsymbol{\theta}_{jm} \sim Dir(\boldsymbol{\alpha})$
  end for
  for all token i in sentence s of doucment j(i=1...N_j) **do**
    Draw $u_{ji} \sim Multi(\boldsymbol{\psi}_{js})$
    Draw $z_{ji} \sim Multi(\boldsymbol{\theta}_{ju_{ji}})$
    Draw $w_{ji} \sim Multi(\boldsymbol{\phi}_{z_{ji}})$
  end for
end for

3. 先行研究

本章では，先行して提案されている Titov らの MG-LDA について説明する．まず，MG-LDA の概要を紹介した後，実際に提案手法と比較する simplfied MG-LDA（sMG-LDA）を導入する．次に，MG-LDA の重要な要素である，スライディングウィンドウ（以下単にウィンドウ）について考察する．

MG-LDA [1] は評価文書からのトピック推定のために提案されたモデルである．MG-LDA は文書単位で推定されるグローバルトピックと，ウィンドウ単位で推定されるローカルトピックの 2 種類のトピックを仮定する．とくに本研究で検討する評価側面はローカルトピックと対応付けられる．本稿では，ローカルトピックのみに着目することから，グローバルトピックはモデルから除外し，ローカルトピックのみを仮定するモデルである sMG-LDA を導入し，提案手法との比較に利用する．

3.1 sMG-LDA

sMG-LDA の生成過程を Algorithm 1 に示す．この生成過程を説明する前に，本稿で用いる表記について説明する．まず， T はトピック数，D は文書数，S_j は文書 j 内の文数をそれぞれ表す．また M_j は文書 j のウィンドウ数を表す．ウィンドウ数はウィンドウサイズ K と文数 S_j によって決定され，$M_j = K + S_j - 1$ である．さらに，q〜$Dir()$ はディリクレ分布，q〜$Multi()$ は多項分布から生成されることをそれぞれ表す．加えて $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$，$\boldsymbol{\gamma}$ はそれぞれディリクレ分布のパラメータベクトルであり，モデルのハイパーパラメータである．

まず，sMG-LDA はディリクレ事前分布 $Dir(\boldsymbol{\beta})$ からトピック t ごとに単語の生成確率ベクトル $\boldsymbol{\phi}_t$ を生成する．次に $Dir(\boldsymbol{\gamma})$ から各文書 j の文 s ごとにウィンドウの生成確率ベクトル $\boldsymbol{\psi}_{js}$ を，$Dir(\boldsymbol{\alpha})$ からウィンドウ m ごとにトピックの生成確率ベクトル $\boldsymbol{\theta}_{jm}$ をそれぞれ生成する．最後にこれらの確率ベクトルか

ら次の 3 種類の変数を生成する．u_{ji} は文書 j の単語トークン i である，トークン ji のウィンドウの割り当て，z_{ji} はトピックの割り当て，w_{ji} は単語の割り当てをそれぞれ表す．これらの変数は各々対応する多項分布から生成される．なおウィンドウサイズが 1 のとき，このモデルは文ごとに Bag-of-words を仮定した LDA と等価である．

本研究では，sMG-LDA のパラメータ推定を collapsed Gibbs Sampling によって行う．推定のためには，各トークンのトピックとウィンドウの割り当てである，z と u をそれぞれ条件付き分布からサンプリングする必要がある．トークン ji での各潜在変数の条件付き分布の式は以下で表される．

$$p(z_{ji}=t|w_{ji}=v,u_{ji}=m,\mathbf{w}^{-ji},\mathbf{z}^{-ji},\mathbf{u}^{-ji}) = \frac{n^{v}_{t,-ji}+\beta}{n^{v}_{\cdot,-ji}+V\beta}\frac{n^{j,m}_{t,-ji}+\alpha_t}{n^{j,m}_{\cdot,-ji}+\alpha_0} \quad (1)$$

$$p(u_{ji}=m|z_{ji}=t,\mathbf{w},\mathbf{z}^{-ji},\mathbf{u}^{-ji}) = \frac{n^{j,m}_{t,-ji}+\alpha_t}{n^{j,m}_{\cdot,-ji}+\alpha_0}\frac{n^{j,s}_{m,-ji}+\gamma_m}{n^{j,s}_{\cdot,-ji}+\gamma_0} \quad (2)$$

ここで $n^{v}_{t,-ji}$ はトークン ji を除くトピック t で単語 v が出現した回数を，$n^{j,m}_{t,-ji}$ はトークン ji を除く文書 j のウィンドウ m でトピック t が出現した回数を，$n^{j,s}_{m,-ji}$ はトークン ji を除く文書 j の文 s でウィンドウ m の割り当てられた回数をそれぞれ表す．またドットはその変数のカウントを合計したカウントを表す．

ハイパーパラメータベクトル $\boldsymbol{\alpha}$ と $\boldsymbol{\gamma}$ については Minka が提案する定点反復法 [4] により推定する．つまり $\boldsymbol{\alpha}$ と $\boldsymbol{\gamma}$ の各要素 $\alpha_t(t=1...T)$，$\gamma_m(m=1...K)$ は推定によって求めた．なお $\alpha_0=\sum_{t=1}^{T}\alpha_t$，$\gamma_0=\sum_{m=1}^{K}\gamma_m$ である．残る $\boldsymbol{\beta}$ については全て同じ値で固定した．具体的には $\boldsymbol{\beta}$ の各要素 β は $\beta=0.1$ と設定する．

3.2 先行研究の考察

ここで，MG-LDA のウィンドウについて考察する．実際のウィンドウの例を図 1 の左側に示す．ウィンドウはローカルトピックを考慮するため，いくつかの文を覆うように文書中で定義される．図 1 はウィンドウサイズが 3 の例であり，サイズが大きいほど，より広い範囲の文を考慮できる．また，単語ごとにウィンドウの割り当てを仮定することで，ウィンドウの重なりを許すようなモデル化が可能である．図のように，ウィンドウは文書内に網羅的に用意され，各単語はパラメータ推定時にどのウィンドウに属するか他の変数と合わせて推定される．

このウィンドウを利用したモデル化には，いくつかの制約がある．まず，ウィンドウサイズは対象の文書集合全体で固定する必要がある．実際の文書には長い文書や，短い文書が混在しているため，1 文からなる文書に対しても，大きいウィンドウサイズが割り当てられてしまう場合がある．また，図 1 のようにウィンドウは網羅的に用意されるため，実際にはほとんど使用されないウィンドウも多いと考えられる．これらは冗長的な表現とみられ，モデルの精度に影響している可能性がある．

そこで，図 1 の右側のように実際の文書中のトピックを反映した，簡潔な範囲によって表現することを検討する．必要な数だけウィンドウを代替する要素を用意し，各々の範囲ごとにローカルトピックを推定する．これにより，文書中の話題の長さに応じて，範囲を変化させることができるため，上記で挙げたウィンドウの制約を補うモデル化が可能だと考えられる．次章では，この実現のために dd-CRP を用いた新しいモデルを提案する．

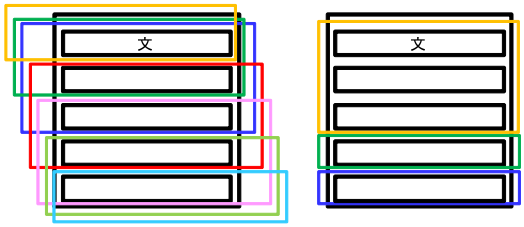

図 1　ウィンドウの網羅的なモデルと提案手法で目指す簡潔なモデル

4. 提 案 手 法

本章では，新たに dd-CRP を応用した，評価側面を推定するためのモデルを提案する．はじめに提案手法に利用する dd-CRP の概要を説明する．次に，提案モデルを説明し，最後に collapsed Gibbs Sampling を用いたパラメータ推定を記す．

4.1 dd-CRP

dd-CRP は Blei らによって提案された分割上の確率分布である [2]．dd-CRP の特徴は，通常の Chinese restaurant process（CRP）[5] にない変数間の距離的な依存関係を考慮できる点である．CRP と dd-CRP は様々な問題へ適用することが検討できるが，ここではデータを確率的にクラスタに割り当てる場合を通して説明する．

まず，CRP について簡単に説明する．各データを表す確率変数をクラスタに割り当てることを考える．既に $i-1$ 番目までのデータがクラスタに割り当てられたとき，i 番目のデータをクラスタに割り当てる確率は以下の式で表せる．

$$p(x_i=k|X^{i-1},\gamma)\propto\begin{cases} n_k & (\text{既存クラスタ}) \\ \gamma & (\text{新規クラスタ}) \end{cases} \quad (3)$$

ここで，$p(x_i=k|X^{i-1},\gamma)$ は $i-1$ 番目までのデータ（X^{i-1} は $i-1$ 番目までの変数の集合）がクラスタに割り当てられたとき，i 番目のデータに対するクラスタの割り当てを表す確率変数 x_i がクラスタ k をとる確率である．n_k はクラスタ k に既に割り当てられているデータの数であり，γ は集中度パラメータである．CRP はディリクレ多項分布の極限として解釈することができる．また新規クラスタから生成される確率をもつため，次元数（クラスタ数）が固定されない．そのためデータやモデルの複雑さに応じた，任意の次元数をもつ確率分布を構成することができ，混合モデルの拡張を中心に応用されている．

次に，dd-CRP について説明する．dd-CRP はデータをクラスタに割り当てるのではなく別のデータへ割り当てる（リンクさせる）ことを考える．データ i をデータ j へリンクさせるとき dd-CRP では次式に従う．

$$p(c_i = j|D,\gamma) \propto \begin{cases} f(d_{ij}) & (i \neq j) \\ \gamma & (i = j) \end{cases} \quad (4)$$

ここで f は減衰関数，d_{ij} は定義される i と j との距離，D は全データの距離の集合，γ は 0 より大きい尺度パラメータをそれぞれ表す．減衰関数 f は距離に従って決まる非増加関数であり，非負の有限値をとり，かつ $f(\infty) = 0$ である．$p(c_i = j|D,\gamma)$ は D と γ が与えられたとき，i 番目のデータがリンクするデータを表す確率変数 c_i が，データ j を値にとる確率である．dd-CRP では，自分以外の他のデータへのリンクは減衰関数に，自分自身へのリンクは γ にそれぞれ従う．この過程からデータ同士のリンク関係が得られる．他のデータへリンクする確率は上述の式の減衰関数によって制御されるため，距離が近いほど結びつきやすいというバイアスをかけることができる．

全てのリンクが与えられたとき，これは間接的にデータのクラスタへの割り当てを表す．c_i が得られたとき，データ i のクラスタへの割り当てを $x(c_i)$ と表す．提案するモデルでは，dd-CRP を周辺の文同士の依存関係をモデル化するために用いる．

一般に，dd-CRP はデータ間の距離のみで決まるが，本稿では，その特別な場合である sequential CRP を利用する．sequential CRP はデータをリンクさせるとき，系列中の自分より後ろに位置するデータへはリンクせず，自分自身か前のデータのみにリンクするモデルである．これより sequential CRP はリンクの循環構造が発生しないため，幾分簡便になる．次節で示す提案モデルでは，同じ文書内の文ごとに dd-CRP から生成する変数を仮定する．これは 1 次元の系列をモデル化することになるため，sequential CRP を採用する．ただ，表記を統一させるため，本節以降でも dd-CRP と記述する．

4.2 モデル概要

本節では，提案モデルの概要を述べる．トピックと単語の生成過程や，文単位での Bag-of-words を仮定する点は MG-LDA と同様であるが，提案モデルではウィンドウに代わる要素として，各文に対して潜在クラスを仮定する．この潜在クラスが dd-CRP から生成されると仮定することで，リンクされた変数間で文書の部分集合を構成する．実際，前節でみた dd-CRP の特性は 3.2 節で検討したモデル化の実現に適していると言える．

提案手法の具体的な生成過程を Algorithm 2 に示す．ここで，q〜dd-CRP(D,f,γ) は距離集合 D，減衰関数 f，尺度パラメータ γ が与えられたもとで，dd-CRP から変数 q が生成されることを表す．また，c は文書の各文に仮定される潜在変数であり，文書 j の文 s では c_{js} と表す．これが dd-CRP によって生成され，文がどの文へリンクするかを表す確率変数になる．この c_{js} は間接的に文書中の文の部分集合に対するクラスタの割り当てを与える．これを $x(c_{js})$ と表す．残りの表記については sMG-LDA および前節の dd-CRP の表記をそのまま継承する．

まず，提案モデルは $Dir(\boldsymbol{\beta})$ からトピックごとに単語の生成確率ベクトル $\boldsymbol{\phi}_t$ を生成する．次に各文書の各々の文に対して潜在クラスの割り当て c_{js} を dd-CRP(D,f,γ) から生成する．これにより間接的に得られるクラスタの数だけ，$Dir(\boldsymbol{\alpha})$ からトピックの生成確率ベクトル $\boldsymbol{\theta}_{jk}$ を生成する．最後にこれらの確率ベクトルから sMG-LDA 同様，z_{ji} および w_{ji} をそれぞれ対応する多項分布から生成する．

Algorithm 2 Generative process of Proposed Model

for all topic t (=1...T) **do**
  Draw $\boldsymbol{\phi}_t \sim Dir(\boldsymbol{\beta})$
end for
for all document j (=1...D) **do**
  for all sentence s of document j (s=1... S_j) **do**
    Draw c_{js} ∼dd-CRP(D, f, γ)
  end for
  for all latent class k of document j (k=1...) **do**
    Draw $\boldsymbol{\theta}_{jk} \sim Dir(\boldsymbol{\alpha})$
  end for
  for all token i in sentence s of doucment j(i=1...N_j) **do**
    Draw $z_{ji} \sim Multi(\boldsymbol{\theta}_{jx(c_{js})})$
    Draw $w_{ji} \sim Multi(\boldsymbol{\phi}_{z_{ji}})$
  end for
end for

ここで，提案モデルで仮定する距離 D と減衰関数 f を定義する．まず，dd-CRP から生成される変数 c は文ごとに定義されるため，その変数間の距離は単純に文間の差分とする．また，この距離は離散値をとることから，減衰関数 f には窓関数を採用する．閾値 a と変数 i，j 間の距離 d_{ij} のもとで窓関数 f は以下のように定義する．

$$f(d_{ij}) = \begin{cases} 1 & (d_{ij} \leqq a) \\ 0 & (d_{ij} > a) \end{cases} \quad (5)$$

閾値の値が大きければ，より遠くの文とリンクするようになるが，特に閾値を $a = 1$ と設定すれば，隣接する文同士のみリンクされる．

提案モデルの概要図を図 2 に示す．図では窓関数の閾値を $a = 1$ として，同一の潜在クラスが連続になる場合を示している．次章の実験でもこの閾値 $a = 1$ のモデルで評価する．このとき潜在クラスは図 1 の右側のように，文書を直和分割するように割り当てられる．

繰り返しになるが，dd-CRP から生成される潜在変数 c は間接的に潜在クラスの割り当て x を与える．この潜在クラスごとにトピックの分布を表す確率ベクトル $\boldsymbol{\theta}$ の存在を仮定する．図 2 では 1 文書内の 3 つの潜在クラスがある箇所の例を示している．各クラスに対して $\boldsymbol{\theta}$ が仮定され，文内のトピック z はここからサンプルされる．単純な文単位でトピック分布を仮定する場合と異なり，同様の話題が書かれた周辺の文で，一つのトピック分布を共有する．また，dd-CRP を用いたことでデータの複雑さに応じて任意のクラス数に分割でき，かつ周辺の文ほど同じ文脈の上で記述されているというバイアスも考慮できる．特に 1 つ目の特徴はウィンドウサイズを文書集合全体で固定しなければならない sMG-LDA と異なり，文書の長さに応じて考慮する文の範囲を適応的に変化できる．

sMG-LDA では，ウィンドウ（提案手法の潜在クラスと対応する）の割り当ては各トークンに対して割り当てる．そのため，

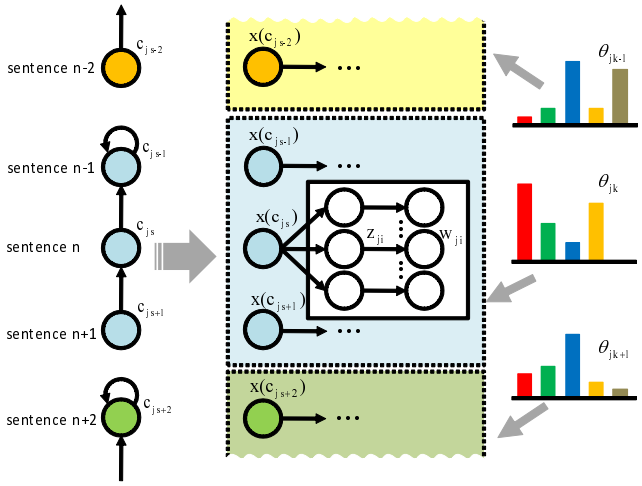

図 2　提案モデル概要図

図 1 の左側のように，文ごとに重なりを許すようなモデル化が可能である．対して提案手法では，先ほど述べたように直和分割するようにモデル化する．そのため，提案手法が MG-LDA を完全に包含しているモデルとはなっていない．

4.3　パラメータ推定

提案モデルを collapsed Gibbs Sampling により推論する．パラメータを推定するために，トピックの割り当てである潜在変数 z と dd-CRP から生成される潜在変数 c の 2 つをサンプリングする必要がある．まず，z に関する条件付き分布は以下のように定式化できる．

$$p(z_{ji}=t|w_{ji}=v, x(c_{js})=k, \mathbf{w}^{-ji}, \mathbf{z}^{-ji}, \mathbf{x}(\mathbf{c}^{-js})) = \frac{n^{v}_{t,-ji}+\beta}{n^{v}_{\cdot,-ji}+V\beta}\frac{n^{j,k}_{t,-ji}+\alpha_t}{n^{j,k}_{\cdot,-ji}+\alpha_0} \quad (6)$$

式（6）は sMG-LDA の式（1）に対応し，その右辺第 1 項は同じである．第 2 項の $n^{j,k}_{t,-ji}$ はトークン ji を除く文書 j の潜在クラス k でトピック t が出現した回数を表す．

次に，c に関する条件付き分布を示す．文書 j の文 s の潜在変数 c_{js} が同じ文書の文 r の変数へリンクするとき，その条件付き分布は以下のように表される．

$$p(c_{js}=r|\mathbf{w},\mathbf{z},\mathbf{c}^{-js},\boldsymbol{\alpha}) \propto \begin{cases} \gamma & (s=r) \\ f(d_{sr})q(\mathbf{z},\mathbf{x},\boldsymbol{\alpha}) & (s\neq r) \end{cases} \quad (7)$$

ここで，

$$\begin{aligned} q(\mathbf{z},\mathbf{x},\boldsymbol{\alpha}) &= \frac{p(\mathbf{z}_{\mathbf{x}^k(\mathbf{c}^{-js})\cup\mathbf{x}^l(\mathbf{c}^{-js})}|\boldsymbol{\alpha})}{p(\mathbf{z}_{\mathbf{x}^k(\mathbf{c}^{-js})}|\boldsymbol{\alpha})p(\mathbf{z}_{\mathbf{x}^l(\mathbf{c}^{-js})}|\boldsymbol{\alpha})} \\ &= \frac{\prod_{t=1}^{T}\Gamma(\alpha_t)}{\Gamma(\alpha_0)}\frac{\Gamma(n^{j,k}_{\cdot}+\alpha_0)}{\prod_{t=1}^{T}\Gamma(n^{j,k}_{t}+\alpha_t)} \\ &\quad * \frac{\Gamma(n^{j,l}_{\cdot}+\alpha_0)}{\prod_{t=1}^{T}\Gamma(n^{j,l}_{t}+\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(n^{j,k+l}_{t}+\alpha_t)}{\Gamma(n^{j,k+l}_{\cdot}+\alpha_0)} \end{aligned} \quad (8)$$

である．$n^{j,k}_{t}, n^{j,l}_{t}$ はそれぞれ文書 j の潜在クラス k（l）でトピック t が出現した回数を表し，$n^{j,k+l}_{t}$ はこの 2 つのカウントの和である．また $\Gamma()$ はガンマ関数である．また，$\mathbf{z}_{\mathbf{x}^k(\mathbf{c}^{-js})}$ は c_{js} を除く潜在クラス k に割り当てられている文に属するトークンの位置にある z の変数集合であり，$\mathbf{z}_{\mathbf{x}^k(\mathbf{c}^{-js})\cup\mathbf{x}^l(\mathbf{c}^{-js})}$ は潜在クラス k と l の上記 z の変数集合を合わせた集合である．

式（7）の第 1 段は自分自身へリンクする場合，第 2 段は他変数へのリンクする場合である．また式 (8) は式 (7) 第 2 段の第 2 項であり尤度関数にあたる部分である．別々だった潜在クラスが結合するときのみ尤度の変化が起こるため，この項は潜在クラスを結合する前（分母）と結合した後（分子）での尤度比を表している．

ハイパーパラメータ γ は Blei ら [2] が提示している Griddy Gibbs method [6] を利用して推定する．これは条件付き分布の推定が困難であるとき，いくつかの離散格子点での値を計算し，これら離散点からサンプリングする方法である．このためやや厳密性に欠ける手法ではあるが，格子点上での推定値を算出できる．次章での実験では，0.005 刻みで 0.005 から 5.0 までの値を格子点として推定している．

5.　実　　験

本章では，定量的な評価結果を示す．3.1 節で導入した sMG-LDA のウィンドウサイズを，1 から 5 まで変化させた 5 つのモデルと，提案手法である dd-CRP を応用したモデルの計 6 モデルを評価する．

以降で実験の概要を説明する．まず，デジタルカメラに関する日本語と英語のデータセットをそれぞれ用意した．日本語データは価格.com [7] から 13,638 件，英語データは Amazon.com [8] から 11,279 件分をそれぞれ取得した．日本語データについては，特徴語として名詞を利用し，高頻度語は除外した．英語データについては，頻度 1 の単語と一般語を除去し，ステミング処理を行なっている．このような条件のもとで得られた各データセットの語彙数は日本語データが 11,577 語，英語データが 8,487 語である．

次に，評価指標であるが本稿では，パープレキシティにより評価する．パープレキシティは言語モデルの評価などに用いられる指標であり，新たな単語が与えられたとき，学習したモデルからどれだけ予測単語を絞り込めているかを示す．パープレキシティはモデルから得られる予測分布を用いて以下のように与えられる．

$$\exp\left(-\frac{1}{N^{test}}\log p(\mathbf{w}_{test}|\mathbf{w}_{training})\right) \quad (9)$$

ここで N^{test} はテストデータ用のトークン数，$\mathbf{w}_{test}$ はテストデータのトークン集合，$\mathbf{w}_{training}$ は学習データのトークン集合である．パープレキシティが小さければ，小さいほど予測時のモデルの複雑さが小さく抑えられており，高い精度をもつモデルと考えられる．本稿では，評価文書内のトークン総数の 90%を学習データ，10%をテストデータとして利用する．ただし，提案モデルで利用する dd-CRP は予測分布を推定する際に，未決定の変数については Gibbs Sampling で再学習して推定する必要がある．これは非常に学習に時間がかかるため，今回は各文内のトークンが全てテストデータに割り当てられない

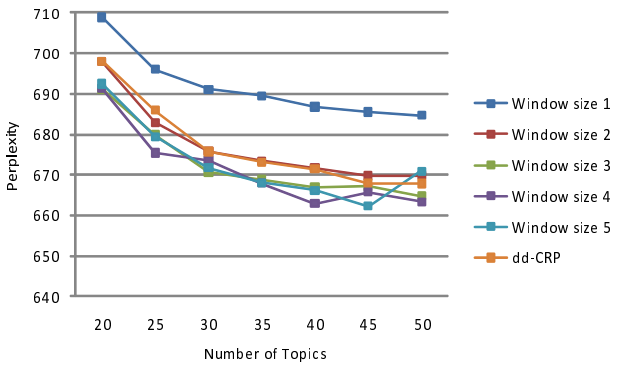

図 3　日本語データに対するパープレキシティの比較

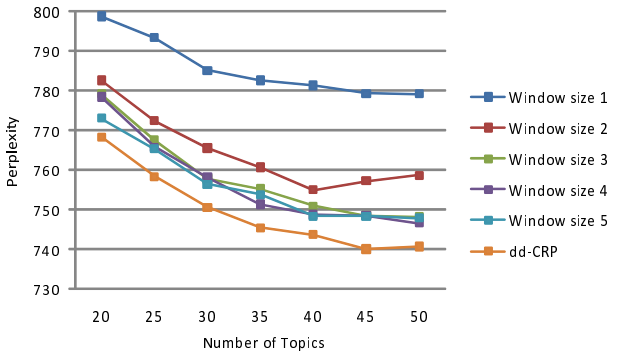

図 4　英語データに対するパープレキシティの比較

ように学習データとテストデータを分割する．トピック数は 20 から 5 トピックずつ増やし 50 まで算出する．

日本語データの結果を図 3 に，英語データの結果を図 4 にそれぞれ示す．まず，比較する sMG-LDA の振る舞いを見る．どちらのデータセットでも，ウィンドウサイズの増加とともに，パープレキシティが低下している．これはより広い範囲を考慮できるためであり，改善していることがわかる．ウィンドウ 3 以降ではパープレキシティの減少の幅が小さくなり，ほぼ差がなくなっているため，良いモデル選択がサイズ 3 から 5 程度であることが分かる．

次に，提案手法と比較する．日本語データでは，sMG-LDA のウィンドウサイズが 3 から 5 のモデルが低いパープレキシティを示している．これに対して，英語データでは提案手法が全体的に最も低いパープレキシティを示している．このため，英語データに関しては提案手法が有効に働いていることがわかる．

これらの違いは今回使用したデータの特徴が影響していると考えられる．提案手法の結果が良かった英語データをみると，全体的に文書中のトピックがはっきりとしており，話題の変わり目がはっきりとしている傾向があった．提案手法では文書を直和分割するようにモデル化しているため，この仮定が英語データに対しては上手く働いたのではないかと考えられる．対して日本語データでは，ある観点について評価している途中で，その観点に関係のない情緒的な文が挟まれる傾向が見られた．提案手法では必ず同じ潜在クラスが隣接するようにモデル化されるため，ある評価側面に関する記述の中で，文書中に情緒的

表 1　提案手法によるトピックの例 1（トピック数 30）

トピック 1		トピック 2	
単語	確率値	単語	確率値
デザイン	0.1650	バッテリー	0.1414
感	0.0382	枚	0.0439
質感	0.0260	撮影	0.0327
感じ	0.0257	予備	0.0322
好き	0.0220	充電	0.0263
高級	0.0206	電池	0.0262
色	0.0206	日	0.0245
購入	0.0159	一	0.0187
シンプル	0.0129	旅行	0.0156
黒	0.0128	使用	0.0154

な記述があると，dd-CRP のリンクが起こらない場合がある．その点ウィンドウは重なりを許すことで，情緒的な文を超えてモデル化ができるため，そのような問題に強いのと考えられる．このような理由から，日本語データでは期待したような改善が見られなかったのではないかと推察される．

英語データで改善がみられているため，提案手法のような必要な数だけ潜在クラスを仮定するモデル化はある程度有効だと考えられる．これに加えて，提案手法をウィンドウのように，重なりを許すようなモデル化が可能であれば，日本語データに対しても改善が見込めると考えられる．

6. おわりに

本稿では，評価文書から評価側面を推定する新たなトピックモデルを提案した．パープレキシティを用いた定量的な評価によって先行研究との比較を行い，とくに英語データセットについては，先行研究を上回る性能を示し，その有効性を確認した．

本稿では，パープレキシティによる既存手法との定量的な比較のみ行った．今後はトピックモデルによって評価側面をどの程度推定できているか検証する必要がある．そのためには，実データに対して何らかのタスクを設け，評価を行う必要がある．これについては今後の課題とし，ここでは，実際に取得できたトピックについて簡単に考察する．

5 章の日本語のデジタルカメラの評価文書に対するトピックの推定結果の一部を表 1 に示す．トピック数は 30 で学習し，そのうち 1 章で提起した評価側面である「デザイン」「バッテリー」と対応付けられる 2 つのトピックを示した．2 つの評価側面を指す単語が出現しており，かつそれらと関連ある単語が上位に来ている．このように想定したトピックの推定にある程度成功しているといえる．

次に，同様の評価文書に対して推定された別のトピックを表 2 に示す．同じく 2 つのトピックを示したが，これらは「購入」「比較」といった単語が並び，必ずしもデジタルカメラを評価する際の観点・基準とは呼べないトピックである．学習に利用した文書を見ると「○○を購入しました。」,「○○を△△と比較してレビューします。」といった記述が多く現れていた．本稿では文書中の記述全てを使用しているため，評価文書中の話題としては理解できるが，必ずしも想定した評価側面とは呼べ

表 2　提案手法によるトピックの例 2（トピック数 30）

トピック 1		トピック 2	
単語	確率値	単語	確率値
購入	0.0614	円	0.0933
機種	0.0337	購入	0.0725
比較	0.0197	万	0.0561
こちら	0.0179	価格	0.0347
レビュー	0.0168	値段	0.0147
参考	0.0161	2	0.0144
他	0.0142	店	0.0131
評価	0.0134	キタムラ	0.0123
私	0.0123	1	0.0112
店頭	0.0117	台	0.0111

ないトピックも推定される結果となった．なお，このようなトピックは提案手法，先行研究ともに現れる．

このようにトピックモデルの適用により，評価文書中からある程度評価側面と呼べるトピックを推定できている一方，全てのトピックが何らかの評価の軸を表すものではないことを確認した．これらを何らかの形で区別するように考慮できる方法の検討が必要といえる．

提案手法に関する課題として，5 章でも述べたように，ウィンドウと同様に潜在クラスの重なりを許すようなモデル化を実現することが挙げられる．これには，例えば，単語単位で dd-CRP から生成される潜在変数を導入することが考えられる．これにより，文ごとに複数の潜在クラスをもつことができるため，重なりを考慮しつつも，必要な数の潜在クラスのみでモデル化が可能だと考えられる．

文　　献

[1] I. Titov, R. McDonald, “Modeling Online Reviews with Multi-grain Topic Models”, *Proc of 17th Word Wide Web Conference*, 2008.

[2] D. Blei, P. Frazier, “Distance Dependent Chinese Restaurant Processes”, *Journal of Machine Learning Research*, Vol.12, pp. 2461-2488, 2011.

[3] D. Blei, A. Ng, M. Jordan, “Latent Dirichlet allocation”, *Journal of Machine Learning Research*, Vol.3, pp.993-1002, 2003.

[4] T. P. Minka, “Estimating a Dirichlet distribution”, Technical report, `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`, 2003

[5] D. Aldous, “Exchangeability and Related Topics”, *École d’été de Probabilités de Saint-Flour XIII*,pp.1-198, 1983.

[6] C. Ritter, M. Tanner, “Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler”, *Journal of the American Statistical Association*, Vol.87, pp.861-868, 1992.

[7] 価格.com, `http://kakaku.com/`

[8] Amazon.com, `http://www.amazon.com/`