統計的言語特性を考慮した評判情報のトピックモデリング

小西　卓哉† 　手塚　太郎†† 　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.ritsumei.ac.jp
††{tezuka,amaeda}@media.ritsumei.ac.jp

あらまし 近年 Web サービスを通じて提供される評判情報が個人レベルから活発に発信されている．本研究ではこれら評判情報の文書データに対してトピックモデルによる解析手法を提案する．既存手法の一つに Titov らによる Multi Grain-LDA があるが，本稿では文の連結による 2 段階の学習をも用いたその性能向上を提案する．さらに Pitman-Yor トピックモデルを応用することで，さらなる精度向上が可能か検討する．

キーワード 評判情報，トピックモデル，テキストモデリング

Review Topic Modeling with Statistical Language Property

Takuya KONISHI† 　Taro TEZUKA†† 　Fuminori KIMURA††† 　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.ritsumei.ac.jp
††{tezuka,amaeda}@media.ritsumei.ac.jp

Abstract In recent years, online review provided through Web services are contributed by many users. In this paper, we propose a method to analyze these online review documents by using a topic model. While Titov and McDonald suggested Multi-grain LDA for this task, we propose its performance improvement by 2 level learning model. This model connects sentences using 1st learing results. In addition, we apply Pitman-Yor Topic model's idea. This model models statistical lexical property "Power-Law" in these review documents. We test this model for improving precision.

Keyword review, topic model, text analysis

1. はじめに

近年 Web の発展に伴い個人からの情報発信が容易となっている．特に Blog や Twitter などのソーシャルメディアを通して発信される情報は，様々な対象への意見や評価を含んでいる．このような意見情報は Web の発展以前には取得が困難であった情報であり，その情報を集約することで有益な知識の獲得が可能になると考えられる．

意見情報の中でも商品・サービスを評価した評判情報が Web 上で活発に発信されている．代表的なものとして，Amazon[1]や価格.com[2]のような Consumer Generated Media（CGM）を提供する Web サービスが，商品に対して自由な評価を投稿できるシステムを構築している．これらのサイトでは，評価されている対象が明確であり，かつ特定の商品への定型化された情報が集約されているため有用な知識が眠っていると考えられる．

本研究ではこのような評判情報に付随している文書データ（評判文書）に対してトピックモデルの適用による知識発見を提案する．トピックモデルとは文書データを確率的生成モデルによってモデリングする手法である．文書や単語の背後に存在する潜在的な話題（トピック）を仮定し，コーパス中の文書や単語間の関連性を推定することを実現する[3][4].

本研究ではこれらトピックモデルを利用して，評判文書集合がもつ商品・サービスの潜在的な評価基準の推定手法を提案する．商品・サービスには評価される際にポイントとなる性質や側面があると考えられ，本稿ではこれを評価基準と呼ぶこととする．例えばカメラの評判情報には，"カメラのデザイン""撮影した時の画質""操作のしやすさ"といったカメラの良し悪しを判断する評価基準が考えられる．このような評価基準は評判文書内に現れ，文書集合からこれらを推定す

ることは評判情報から知識を得る上で重要な要素だと考えられる．

評判文書のトピックモデルとして Titov らが Multi Grain-LDA（MG-LDA）を提案している[5]．これは本研究で所望する評価基準を推定する手法を提供している．本研究ではこの MG-LDA に対して，文の連結を用いた 2 段階学習による予測精度の向上を提案する．また佐藤ら[6]によって提案されている Pitman-Yor トピックモデルを応用することによって，さらなる性能向上が可能か検証する．

2 章で本稿において扱う評判文書を解説し，3 章では先行研究について紹介する．4 章では提案手法について述べ，5 章では評価実験と本手法の応用例について示す．最後に 6 章で今後の課題と展望について議論する．

2. 本研究で扱う評判文書

トピックモデルによる文書のモデル化を検討する上で，評判文書について説明する必要がある．本章では本稿で扱う評判文書について紹介する．

まず本稿で扱う評判文書集合とは，ある 1 つの商品カテゴリについて記述された評判文書の集合と仮定する．商品カテゴリとは例えば“パソコン”“携帯電話”“テレビ”のような大域的な商品のカテゴリである．評判文書集合は単一の商品カテゴリに属する，個別の商品について評価した評判文書から構成される．パソコンの評判文書集合であれば，様々なパソコンが評価された文書を集まり，一つのコーパスを構成する．このような特定の商品カテゴリの評判文書集合を利用するため，一般的なトピックモデルの実験等で利用される新聞記事などと比較すると，相対的に狭い話題が展開される文書集合を解析対象とする．

本稿では評判文書集合を価格.com が提供する商品カテゴリ「デジタルカメラ」から 13638 件分取得し，その中から名詞のみを取り出して素性データとした．図 1 は価格.com で実際に取得したデジタルカメラの評判文書の一例である．この文書では一重線を引いたセンテンスがカメラの“デザイン”の評価，点線が“画質”の評価，二重線がカメラの“機能性の良さ”の評価をそれぞれ記述している．このように評判文書内には，その対象が評価される軸を表す評価基準が存在していることがわかる．本研究ではこの評価基準を，評判文書集合の学習により推定することを目指す．

3. 先行研究

本章ではまずトピックモデルの代表的な手法である latent Dirichlet allocation（LDA）について簡単に紹介する．次に本稿の提案手法に先立って提案されている 2 つの先行研究について述べる．

この機種はデザインが最高だと思います。画質はコンデジとして普通です。現在所有しているパナソニックやニコンのコンデジと比べて画質でのアドバンテージはほとんど分りません。所有しているデジイチやソニーの NEX と比べると画質では劣ります。センサーというよりレンズの差がでます。特にぼけの表現は無理です。ただし携帯のカメラと比べれば圧倒的に優れていることから何処でも持って行けるコンデジの特性を考えれば大満足です。顔認証は我が家の柴犬も認証します。様々な機能は記念撮影に優れていると思います。露出もやや明るめで記念撮影向けです。絵作りも人物の描写に特化しているとさえ感じます。
機能はいろいろ設定できますが、呼び出すのに数アクション必要ですこし面倒くさいです。基本的にフルオートで撮るカメラだと思います。一言で感想を言うと、機能的にも誰にでも勧められるコンデジらしいコンデジです。

図 1：デジタルカメラの評判文書の一例

3.1 LDA

LDA[4]は確率的生成モデルを用いた文書モデル化手法であり，代表的なトピックモデルとして盛んに研究が行われている．

ここで本稿における LDA のトピックモデルの表記法を示す．T はトピックの種類数，D は文書数，N_j は文書 j におけるトークン数をそれぞれ示す．$\phi_{(t)}$はトピック t における単語の出現確率を表すベクトル，θ_j は文書 j におけるトピックの出現確率を表すベクトルをそれぞれ表し，$w_{j,i}$ は文書 j における i 番目に出現したトークンを，$z_{j,i}$ は文書 j における i 番目のトークンに割り当てられたトピックをそれぞれ表す．またαおよびβはディリクレ分布のパラメータを表す．その生成過程は以下のようにモデル化される．

1. $\varphi_{(t)} \sim \mathrm{Dir}(\beta)$ for t=1…T
2. $\theta_j \sim \mathrm{Dir}(\alpha)$ for j=1…D
3. $z_{j,i} \sim \mathrm{Multi}(\theta_j)$ and $w_{j,i} \sim \mathrm{Multi}(\varphi_{z_{j,i}})$ for i=1…N_j

なお p~Dir(q)および p~Multi(q)は q をパラメータとするディリクレ分布と多項分布から確率変数 p を生成することを表す．このような生成過程を通して単語が生成されることを仮定する．これは LDA 登場以前に提案された PLSI[3]と比較して，点推定でなくベイズ推定である点，事前分布としてディリクレ分布を仮定することで，より自然なスムージングを実現している点からロバストなモデル化が成されている．LDA の潜在

変数（トピック）は近似推論法[4]やサンプリング[7]によって推定する．

LDA をはじめ多くのトピックモデルは Bag-of-words と呼ばれる文書表現を想定したモデルである．これは文書を 1 つの袋（Bag）とみなし，その中に単語（word）が詰められていることを表したものであり，文書内部の語順を無視して単語の出現頻度のみを用いる．単語の出現順序を考慮しないことから，言語モデルの観点からはユニグラムモデルとみなすことができる．

3.2 MG-LDA

評判文書へのトピックモデル適用は Titov らが MG-LDA を提案している[5]．2 章で示したように評判文書は通常の LDA が想定する文書集合と比較して非常に狭い集合をモデル化の対象とする．このような評判文書を対象としてトピックモデルの学習を行う場合，複数の文書内に同様の単語が出現する．例えばカメラの評判文書の場合“画質”という単語はカメラの画質性能を指すため，評価基準を推定する上で重要な単語であるが，カメラの評判文書の多くに出現することが予想でき，特徴量としての抽出が難しい．これは前述の Bag-of-words において文書を 1 つの Bag とみなすことに起因する．このように今回の研究目的である評価基準をトピックとして推定するためには，LDA のようなオーソドックスなモデルを工夫する必要がある．

そこで MG-LDA ではウィンドウと呼ばれる潜在要素を導入する．文書の内部の隣接センテンスを 1 つの集合とみなすことで，通常のトピックモデルでは困難な文書内部の局所的なトピック（評価基準）を推定する．これにより通常の LDA で想定する文書レベルでのグローバルなトピックに加え，ウィンドウレベルでのローカルなトピックの推定が可能となる．この局所的に表れるローカルなトピックこそ評判文書集合の評価基準となる．

ウィンドウの概念を図 2 に示す．ウィンドウは潜在変数として表現され，いくつのセンテンスを覆うか（ウィンドウ幅）はモデル選択の 1 つとして決定する．図 2 はウィンドウ幅が 3 の場合である．この例では中央の「センテンス」に対して 3 つのウィンドウが考えられる．どのウィンドウから生成されるかは確率的に決定されるものとし，学習過程の 1 つに組み込まれている．どのウィンドウから生成されやすいか学習が進むにつれて収束していく．

本稿ではこのウィンドウ単位で生成されるローカルなトピックが評価基準を表すものとし注目する．通常の MG-LDA では文書単位のトピックとウィンドウ単位のローカルトピックの両方を推定するが，本稿ではローカルトピックのみを推定する簡略化した MG-LDA を用いて，次章以降で利用する．

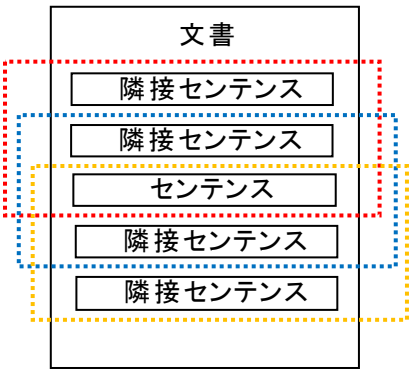


図 2：MG-LDA におけるウィンドウ

3.3 Pitman-Yor トピックモデル

通常の LDA を基に様々なモデルが提案されているが，ここで本研究での応用が期待できる佐藤ら[6]が提案する Pitman-Yor トピックモデル（PYTM）について紹介する．この手法は通常の LDA に加え，文書コーパスに現れる言語特性である Power-Law をモデル化し適応化することで精度向上を実現する．

Power-Law とは冪乗則あるいは冪乗分布と呼ばれる自然界の様々な場面で現れる法則である．自然言語においては単語の出現順位と頻度が反比例の関係になることが，この法則に当てはまるものとして有名である．図 3 は前述のデジタルカメラの評判文書コーパスにおける単語の中で名詞の出現頻度を表したものである．縦軸が単語出現頻度，横軸が各単語の順位を表しており，図 3 はその両対数グラフである．

最も出現頻度の多い“撮影”という単語はコーパス中で 13096 回出現しているが，順位が下がるにつれて出現頻度が反比例的に減少している．ただし低い順位の単語が数多く出現しており，単語分布がロングテールな分布をしていることがわかる．これらの性質はコーパスレベルだけでなく 1 文書の中でも現れる．このような同じ単語が何度も繰り返し出現するという言語の統計的性質は，前述の LDA における単語生成過程に影響を及ぼすと考えられる．

そこで Pitman-Yor トピックモデルでは LDA の生成過程に加えて，Pitman-Yor 過程によって Power-Law の性質に適応化させるモデル化を行っている．近年確率的生成モデルの学習過程においてモデルの複雑さについても学習させる方法として，ノンパラメトリックベイズモデルが注目を集めており，Pitman-Yor 過程もこれを実現する確率過程である．Piman-Yor 過程は Power-Law に従う確率分布を生成することが示されており[8]，これを用いることで同じ単語が何度も繰り返し出現するという自然言語の特性を取り込んだトピ

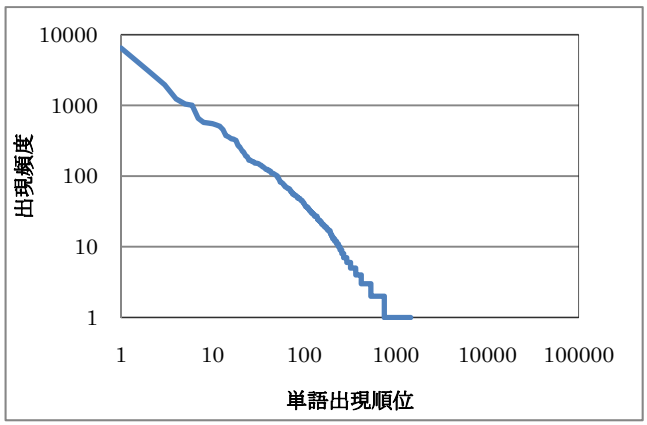


図 3：評判文書コーパスの単語出現頻度および順位

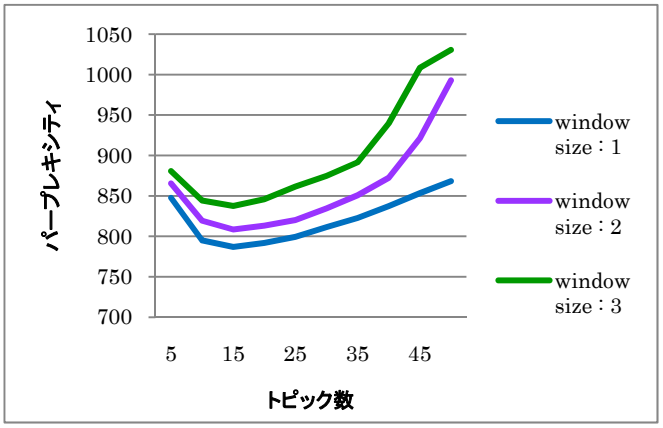


図 4：ウィンドウ幅毎の MG-LDA によるパープレキシティ

ックモデルを構成している．この手法によりトピックモデルのパープレキシティを低く抑えることができ，特に少数のトピック数でのモデル化を行う際にその低下が顕著に表れることが示されている．

以上の 3.2 節および 3.3 節にて紹介した MG-LDA と PYTM を本研究において先行研究とし，本稿における提案手法では，これら 2 つの手法を応用することとする．

4. 提案手法

4.1 従来手法の課題

まず従来手法である MG-LDA の課題について検討する．図 4 は MG-LDA を用いて前述のデジタルカメラの評判文書を学習したモデルに対して，各トピック数におけるパープレキシティをウィンドウ幅毎にグラフ化したものである．ここでパープレキシティとは単語平均予測数を表す指標である．この値が低いほど単語の予測性能が高いと考えられ，精度が高いと言える．図 4 が示すようにウィンドウ幅が 1 のモデルが全体的に低いパープレキシティとなっている．ウィンドウ幅が 1 の場合とは 1 センテンスを 1 つの素性データとして扱う場合である．これはセンテンス毎にトピックを推定することと等価であり，事実上ウィンドウが機能していない．このような結果になる原因は各センテンスが独立に意味を持つ場合が多く，ウィンドウの導入によって冗長性が生まれることで逆に精度が低下するためだと考えられる．ただ，センテンス毎に推定を行うことは 1 つの素性データに含まれる単語が少なくなり，単語の共起性が小さくなる．同一の文脈から生成されたと考えられるセンテンスは 1 つの素性データにまとめて学習を行うことにより，より予測性能の高いモデルを構成することが可能になると考えられる．

4.2 提案手法：2 段階学習

本節では MG-LDA を基に評判文書のトピックモデルを 2 段階に拡張したモデル化を提案する．

ベースとなるのは前章で紹介した MG-LDA である．ウィンドウによって隣接センテンスの集合内でトピック分布を構成することで，通常の LDA では難しい局所的に表れるトピックである評価基準を推定する．しかし 4.1 節で示したようにウィンドウの導入によって，モデルのパープレキシティが低下してしまう．これを防ぐために本稿ではトピックモデルによる学習を 2 段階に分ける手法を提案する．

まず 1 段階目の学習として，各センテンスを 1 つの素性データとした LDA による学習を行う．これは MG-LDA におけるウィンドウ幅 1 のときの学習と等価である．これにより各センテンスに割り当てられるトピック(評価基準)を推定する．

次に 1 段階目の学習により推定されるセンテンス毎のトピックを基に，隣接するセンテンスの連結を行う．隣接するセンテンス間で同じトピックが推定された場合，これらセンテンスが連続した文脈を持っていると仮定する．隣接センテンスが同じトピックを持つ場合，これらセンテンスを連結し，1 つの素性データとして再構成する．連結するセンテンスとそうでないセンテンスが生まれるが，これにより適応的に素性データの集合を構成する．また決定論的なセンテンス集合の構築を行うため，4.1 節で示したウィンドウがもつ冗長性を改善することができると考えられる．

最後にこの新たに生成された学習データに対する LDA による学習を行う．一連の学習過程を図 5 に示す．

1 段階目の段階である程度トピックの推定ができていることを前提とするが，再学習させることにより各素性データがもつ単語の共起情報が増加するため，より高い精度でトピック推定が行えると考えられる．

4.3 PYTM の適用

前節の 2 段階学習に加え，3 章で紹介した PYTM によるさらなる学習精度向上を目指す．PYTM は文書の単語分布を Power-Law に従うような適応化を行うモデル

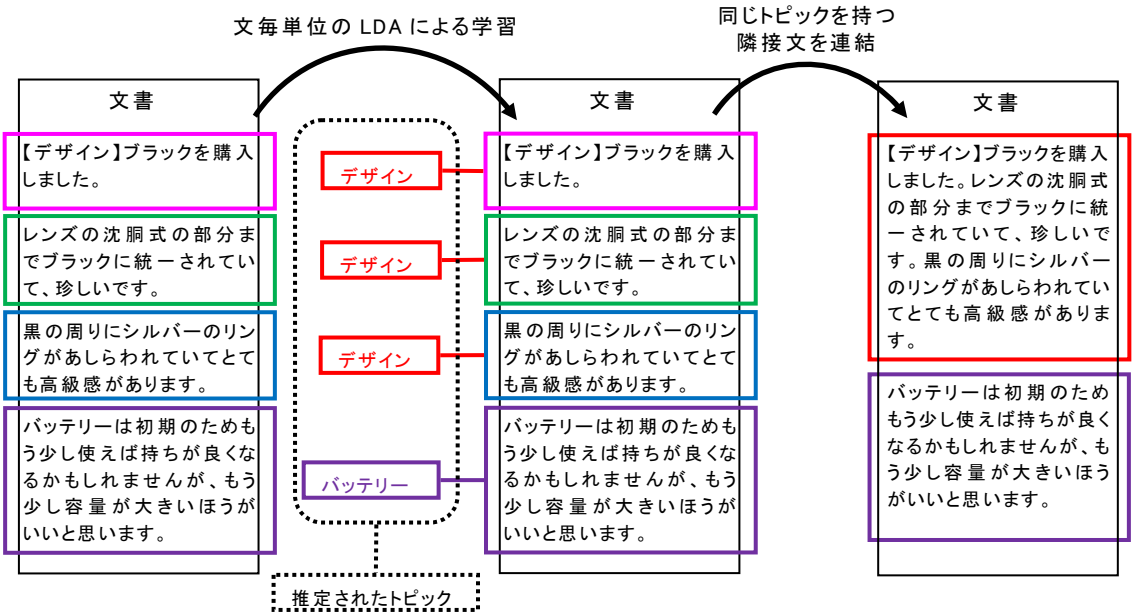


図 5：提案手法．2 段階に分けたセンテンスレベルでの学習過程

として用いられる．本稿では情報量の少ない素性データを補間する目的で本手法を上記の 2 段階手法に対して導入することを検討する．

5. 評価実験および学習結果

本章では提案手法の評価実験結果を示す．また提案手法の学習結果と応用例について紹介する．

5.1 評価実験

本節では提案手法の評価実験結果について示す．評価には 2 章で示したデジタルカメラの評判文書 13638 件から名詞を素性データとするコーパスを用いる．また 4 章で紹介したパープレキシティを評価指標に利用する．今回はデータの 90%を学習データ，10%を訓練データとして学習を行い，3 つの学習サンプルから得られたパープレキシティの平均をとり，従来手法である MG-LDA と提案手法と比較した．さらに 4.3 節で示した PYTM を組み込んだモデルについても同様に評価対象とする．

実際に得られた結果を図 6 に示す． 青色のグラフは従来手法である MG-LDA（ウィンドウ幅:1），緑色のグラフは上記 MG-LDA に対して PYTM を組み込んだモデル，黄色は提案手法である 2 段階学習，赤色のグラフは 2 段階学習にさらに PYTM を組み込んだモデルの結果をそれぞれ示している．

まず青色のグラフと黄色のグラフを比較すると，全体的に従来手法よりも提案手法の方がパープレキシティの低下が確認できる．2 段階に学習を分けることで精度の改善が実現できている．この例から本稿で提案している手法が上手く機能していると言える．

次に PYTM を組み込んだモデル（緑，赤）とそうでないモデル（青，黄色）との比較を行う．若干の違いはあるが，全体的に組み込んだモデルとほぼ同等の精度にとどまっている．このような結果となるのは，モデル内で Power-Law が現れるようなデータがない（もしくはほとんどない）ことが原因だと考察する．素性データ中に前述の Power-Law が出現しない場合，PYTM は通常の LDA と等価な性能を示す．今回は文レベルでの非常に少数の単語を一つの単位とするため，Power-Law が出現しなかったことが，漸近するような結果となった要因だと考えられる．2 段階学習による文の連結によって，1 つの素性中の単語量を増加させることを狙ったが，PYTM において Power-Law を再現する程度の単語量は得られなかったと考察する．

5.2 学習結果と商品特性可視化への応用

本節ではトピックモデルによって得られる評判文書の学習結果について示す．

LDA はトピックの推定によって文書と単語間の関連性を推定する．ここでは各トピックのもとに推定された単語上位 10 語をまとめた結果を表 1 に示す．トピック数は 15 と設定したモデルでの結果である．

表 1 をみると，ある程度意味のある単語が一つのトピックとして抽出できていることが分かる．例えば，トピック 1 はデザインに関する単語が上位に来ている．他にもトピック 8 は携帯性に関する単語，トピック 14 はバッテリーに関する単語がそれぞれ上位に来ている．このようなトピックは本研究で目標とした評価基準を表すトピックを構成している．他のトピックも同じように評価基準となるトピックが推定できているが，中

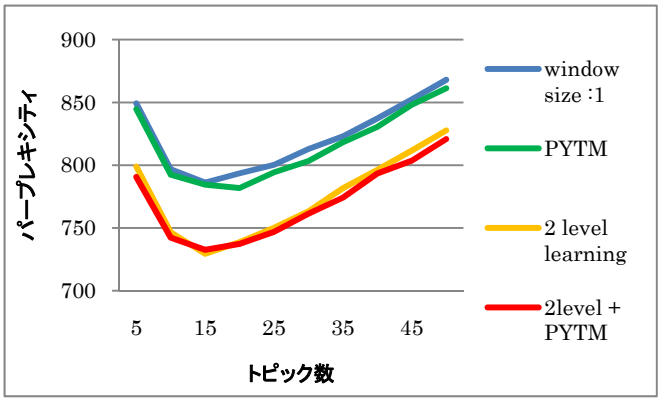


図 6：パープレキシティによる 2 段階学習との比較

には評価トピックとみなすかどうか曖昧なトピックも推定されている．例えば，トピック 0 は数字を表す単語が上位に来ている．このトピックは画素のようなデジタルカメラを定量的に評価しているものと解釈できる一方，トピックとみなすかどうかは意見が分かれる部分だと言える．トピック 5 はカメラの機種やブランドを表している単語が上位に来ている．ブランドというカメラの性質を表す側面だとみなすこともできるが，同じく議論となるものだと考えられる．このように，トピックモデルは教師なし学習によってクラスタリングの一種を実現するものであるため，どのような性質をもったグループ化ができるかはデータ次第である．この点については 6 章でも今後の課題として取り上げる．

次にこれらトピックモデルを用いた応用例について検討する．ここまではコーパス全体からの評価基準推定を行ってきたが，これの個別商品に対する適用を考える．個別の商品はそのカテゴリ中で様々な特性をもつ．例えばカメラであれば「○○という機種はズーム機能が優れている」といったようなものである．ある程度その商品群に詳しいユーザはこのような商品毎の性質を暗黙知として知っていると思われるが，同じ商品について書かれた評判文書群の中には，こうした特性が評価基準の偏りとして出現するのではないかと考えられる．

本稿ではこれまでに導入したトピックモデルを用いて，これら特性の発見を試みる．商品毎の特性を本稿では評価基準の偏りとみなすようにする．もし評判文書の記述中に商品に関する有名な特性が記述されている場合，評価基準の割合が相対的に大きく（あるいは小さく）なると考えられる．

今回はデジタルカメラの評判文書から 100 件以上投稿されている 7 つの商品に対してトピックの偏りを調査した．コーパス全体から推定したトピック分布を用いて各商品に対する評判文書中の各語のトピックを推定し，これを用いて商品毎のトピック分布を算出した．

結果を図 7 に示す．なおトピック数は 18 とした．図 7 の結果からわかるように，他の商品と比べてトピックの分布に偏りのある商品が存在する．例えばトピック 14 では商品 1 と商品 4 が大きな値を示している．このように個別の商品に限定してトピックモデルを応用することで，文書中の記述から商品の大まかな特性を発見することが期待できる．

6. 今後の展望

本稿では MG-LDA の精度向上を目的とした 2 段階学習による評判情報のトピックモデリングについて提案し，実験の結果モデルの性能向上が確認できた．

本研究ではトピックモデルを用いて，評判文書において評価基準を研究対象として扱った．評判情報に関する研究において，この評価基準と同じく重要な研究課題として，評価表現抽出やその極性の判定が挙げられる．これは本稿における提案を応用レベルで検討する上で，重要な要素であり今後の発展が期待される．

また 5.2 節で示したように，推定されるトピックは教師なし学習の結果として得られることから，解析者の意図が介在せずに推定される．これは思いがけない知識の発見が期待できる一方，人間の直感に反するような評価基準を推定してしまうことがある．とくに本稿のように，トピックを特定の意味合いをもつ要素（評価基準）として抽出することを目的とする場合は，ある程度意図的にトピックの方向性の補正を行いたい場合が考えられる．

解決方法の 1 つとしては半教師あり学習が挙げられる．コーパス中にある少数の教師データを混ぜることで，所望するような推定結果を得ることが期待できる．これについては，トピックモデル全般にかかわる課題の 1 つでもあるため，今後の大きな課題といえる．

また本稿で使用した LDA はマルチトピックモデルと呼ばれる手法である．特定のトピックとして点推定するのではなくトピックが出現する割合の推定を実現する．ただ本稿で扱った評判文書ではコーパス全体を通して話題の展開が狭い．このためほとんどの単語・センテンスをユニトピックとして扱うことが検討できる．ユニトピックモデルとしては，混合ディリクレ分布を用いたモデル化が提案されており[9]，検討すべき手法として挙げられる．

最後に本稿で取り上げた評価基準は今後様々な応用方法が考えられる．通常の文書と比較して評判文書は商品や投稿者，地域性といった属性が明示されている．これらと組み合わせることで，評判情報の新たな知識発見のために様々なアプローチができると考えられる．今回例に挙げた商品特性の可視化のように，周辺情報

を用いてコーパスの一部（商品）の特性を発見するといった応用が期待できる．

参 考 文 献

[1] Amazon.co.jp, http://www.amazon.co.jp/

[2] 価格.com，http://www.kakaku.com/

[3] T. Hofmann, “Probabilistic latent semantic indexing”, In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, pp. 50-57. ACM Press, 1999.

[4] D. M. Blei, A. Ng, M. Jordan, “Latent Dirichlet allocation”, Journal of Machine Learning Research, Vol.3, No.5, pp.993-1022, 2003.

[5] I. Titov, R. McDonald, “Modeling Online Reviews with Multi-grain Topic Models”, In Proceedings of 17th International World Wide Web Conference, 2008.

[6] I. Sato, H. Nakagawa, “Topic Models with Power-Law Using Pitman-Yor Process”, In Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining, 2010.

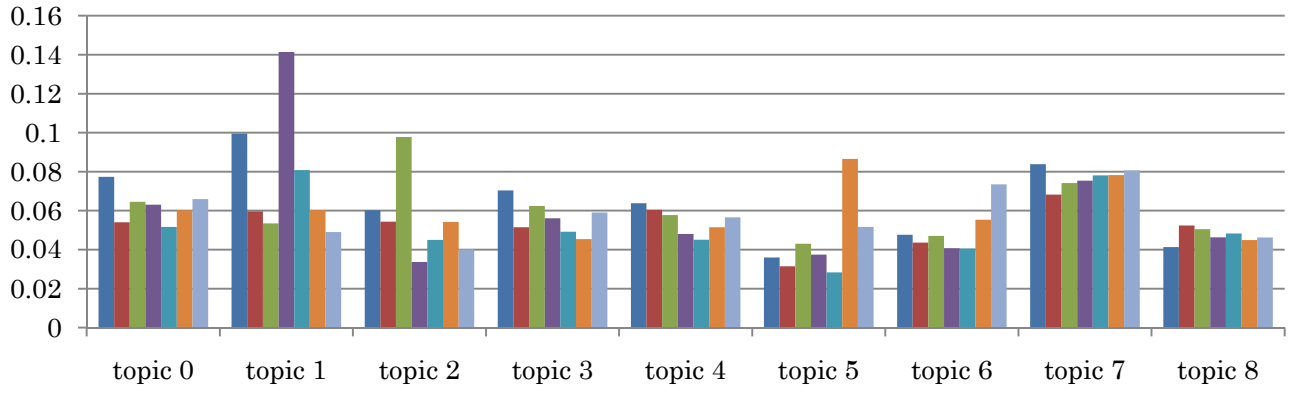
[7] T. L. Griffiths, M. Steyvers, “Finding scientific topics”, In Proceedings of the National Academy of Sciences of the United States of America, Vol. 101, No. suppl.1, pp. 5228-5235, 2004.

[8] S. Goldwater, T. L. Griffiths, M. I. Jordan, “Interpolating Between Types and Tokens by Estimating Power-Law Generators”, In Advances in Neural Information Processing System, 2006.

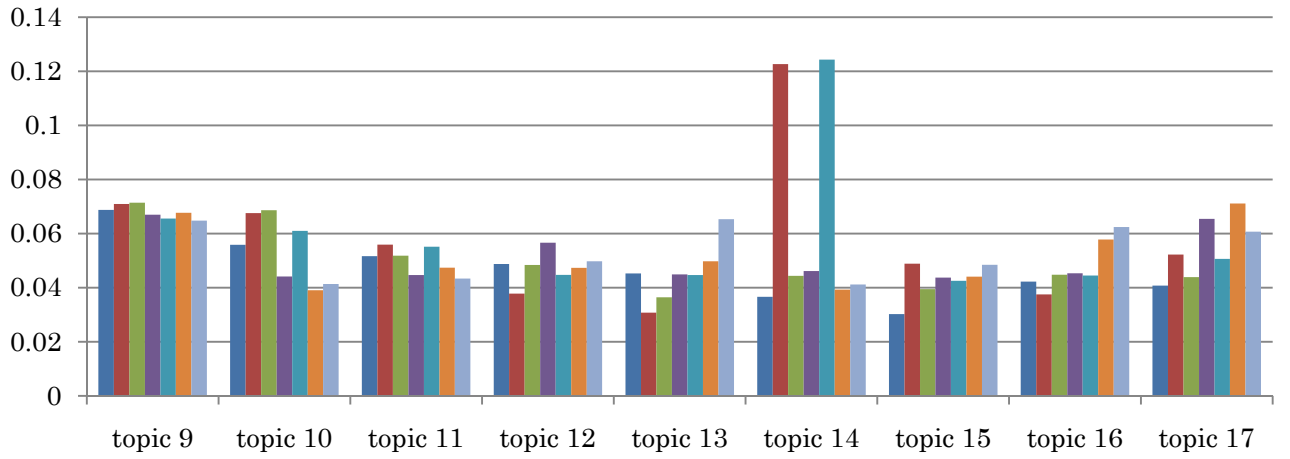
[9] 山本幹雄，貞光九月，三品拓也，“混合ディリクレ分布を用いた文脈のモデル化と言語モデルへの応用”，情報処理学会 研究報告，pp29-34，2003.

topic 0	topic 1	topic 2	topic 3	topic 4
0	デザイン	購入	ホールド感	ズーム
1	感	一	レンズ	倍
2	感じ	年	部分	広角
万	色	前	ケース	レンズ
3	好き	台	ホールド	マクロ
画素	質感	眼	グリップ	光学
5	好み	目	性	望遠
4	個人	レビュー	本体	側
8	発色	価格	問題	倍率
円	高級	使用	指	範囲
topic 5	topic 6	topic 7	topic 8	topic 9
機種	画質	手	携帯性	撮影
機	満足	機能	コンパクト	ノイズ
比較	点	補正	画質	感度
メーカー	評価	モード	性能	フラッシュ
他	度	機能性	機能	室内
IXY	不満	シャッター	サイズ	写真
製品	非常	設定	価格	綺麗
以前	私	オート	十分	夜景
モデル	全体	顔	ポケット	場合
リコー	期待	マニュアル	値段	風景
topic 10	topic 11	topic 12	topic 13	topic 14
写真	液晶	操作性	撮影	バッテリー
私	画像	ボタン	動画	電池
人	綺麗	操作	画	枚
一眼	画面	設定	カード	充電
用	画質	モード	写	使用
自分	きれい	電源	連	日
使用	確認	メニュー	時間	予備
レフ	表示	ダイヤル	静止	一
子供	問題	簡単	音	旅行
初心者	上	シャッター	中	必要

表 2：トピックモデルによる評判情報のトピック(評価基準)推定結果．
各トピックにおいて出現頻度が高いと推定された単語上位 10 語．



topic : 0	topic : 1	topic : 2	topic : 3	topic : 4	topic : 5	topic : 6	topic : 7	topic : 8
機種	撮影	0	写真	購入	レンズ	一	画質	液晶
機	感度	1	人	年	広角	一眼	評価	画面
メーカー	ノイズ	2	私	前	マクロ	使用	私	綺麗
比較	フラッシュ	万	自分	使用	望遠	台	問題	画像
製品	室内	3	顔	円	側	用	期待	表示
モデル	夜景	画素	子供	発売	カバー	レフ	十分	確認
最近	綺麗	5	撮影	買い替え	ボケ	眼	他	きれい
他	場所	4	初心者	店	焦点	機	普通	ファインダー
以前	場合	8	認識	こちら	キャップ	デジタル	レベル	サイズ
シリーズ	中	円	簡単	ヶ月	端	防水	最高	撮影



topic : 9	topic : 10	topic : 11	topic : 12	topic : 13	topic : 14	topic : 15	topic : 16	topic : 17
満足	バッテリー	点	モード	色	手	操作性	デザイン	携帯性
機能	電池	不満	撮影	画質	ズーム	操作	感	ホールド感
度	枚	カード	設定	感じ	動画	ボタン	感じ	コンパクト
価格	充電	残念	シャッター	感	補正	電源	好き	サイズ
十分	日	改善	オート	バランス	倍	メニュー	質感	性
画質	予備	対応	マニュアル	画像	機能	ダイヤル	個人	ポケット
性能	使用	アップ	機能	発色	撮影	設定	高級	ケース
非常	旅行	画像	シーン	絵	画	簡単	ボディ	ホールド
値段	三	以外	写	自然	機能性	再生	シンプル	携帯
総評	必要	ソフト	フォーカス	印象	光学	位置	見た目	グリップ

商品 0	LUMIX DMC-TZ7	商品 4	サイバーショット DSC-HX5V
商品 1	FinePixF31fd	商品 5	LUMIX DMC-LX3
商品 2	IXY DIGITAL 900 IS	商品 6	GR DIGITAL II
商品 3	サイバーショット DSC-WX1		

図 6：デジタルカメラの評判文書集合中の 7 つ商品におけるトピック（評価基準）分布．
縦軸はトピックの出現割合を表す．各グラフ下の表は対応するトピックにおける上位 10 語．
最下表は 7 つの商品の名称．各グラフ左から商品 0 と対応する．