3L3-5

係り受け情報を用いたトピック粒度の細分化に関する検討

A Study on Subdivision of Granularity of Topic using Dependency Information

月岡晋吾 吉川大弘 古橋武
Shingo Tsukioka Tomohiro Yoshikawa Takeshi Furuhashi

名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University

In these days, e-commerce has been expanding with the spread of the Internet. Along with this, reviews for purchasing transaction have been increasing. When a user wants to know goods and services, it takes a lot of time to read all reviews about them. So, quickly understanding outline of the evaluation items is needed. Topic model as represented by Latent Dirichlet Allocation (LDA) can infer topics in documents, without learning data, which can classify evaluation items into topics. However, it is difficult for most of the topic models to get fine-grained topics. When increasing the number of topics, similar topics tend to be generated. In this paper, we propose a method to subdivide topics by the use of dependency information and demonstrate the utility of the proposed method by evaluating the uniformity and the independence for each topic. Moreover, we propose a visualization method of the hierarchical topic structure in the reviews.

1. はじめに

近年，インターネットの普及により，ネットショッピングなどのインターネットを介した購買取引が増加し，それに伴い購買取引に関するレビューの投稿が増加している．また，個人がレビューサイトに旅行などの趣味に関するレビューを投稿する機会も増加している．投稿されたレビューは，企業が自社の商品やサービスに対する評判や特徴を把握する際や，ユーザーが興味のある商品やサービスを調べる際に役立つテキストデータとして注目を集めている．しかし一方で，企業やユーザが，興味を持つレビュー内で評価されている事柄（評価項目）の概略を把握し，興味を持っている内容を詳しく調べるためには，多くの時間が必要となり，容易ではない．

現在，多くのレビューサイトでは，評価項目の概略を把握するという点に対して，有効な対処がなされていない．例えば，Amazon[*1] では，評価項目に関しての記載は無く，価格.com[*2] や Trip.Advisor[*3] では，人手によって付与されたメタデータと，レビュアーによって与えられた評点を利用し，ユーザーにレビュー内容の概略を示しているものの，レビューに付与されているメタデータは人手により数個しか与えられていないため，実際の評価項目を十分には網羅していない．また，人手によりメタデータを付与する際の問題として，評価項目に必要なメタデータがレビューのカテゴリーに依存し，カテゴリー毎に異なる評価項目が必要となるため，メタデータの構築・更新に膨大な労力を要する点や，メタデータとして付与されていないレビュー内の評価項目の扱いが難しいという点が挙げられる．

これらの問題への対処法として，本研究では潜在的ディレクレ配分法 (Latent Dirichlet Allocation: LDA)[Blei 2003] に着目する．トピックモデルをレビュー文書集合に適用することで，レビュー内容を，教師データを用いることなくいくつかのトピックに分類することができ，レビュー内の評価項目を自動で分類・抽出することが可能となる．しかしこのとき，レビューの文書や文に対してトピックの配分を仮定するという一般的な方法では，トピック数を増加させた際に，類似したトピックや，意味の混在したトピックが増加するため，例えば “立地” のトピックに対して，“ホテルまでのアクセス” や，“周辺へのアクセス” など，細かい粒度のトピックを得にくいという問題点がある．そのため，より細かい粒度のトピックを抽出する必要があると考えられる．そこで本稿では，トピックモデルの適用の際に，文の係り受け情報を用いて文を分割することで，文書や文単位での共起範囲よりさらに細かい共起範囲を考慮し，トピック粒度の細分化を行う手法を提案する．さらにトピックを階層構造で可視化する手法を示す．

連絡先: 月岡晋吾，名古屋大学大学院工学研究科，
tsukioka@cmplx.cse.nagoya-u.ac.jp

∗1 http://www.amazon.co.jp/
∗2 http://kakaku.com/
∗3 http://www.tripadvisor.jp/

2. 従来研究

2.1 評価項目の収集

レビュー内の評価項目の概略の把握を目的とした研究では，評価項目の収集や，対象と評価項目と評価値の組みを得る研究が数多く行われてきた．それらの研究では，評価項目の収集に部分的に人手を用いる半自動収集法 [Kobashi 2005] や，人手により大量のラベル付けしたデータを SVM や NB 分類器により学習し，新たな評価項目を得る研究 [Morita 2012] などがある．しかし，学習データを用いる方法では，レビューカテゴリー毎に学習データの構築が必要となる．

また，レビュー内の評価項目の概略の把握を目的とした研究において，潜在的ディレクレ配分法 (LDA) に代表されるトピックモデルを利用する研究も数多くなされてきた．トピックモデルは教師無しで学習を行うため，ドメイン毎に学習データのラベル付けを行う必要がない．トピックモデルを利用し，レビュー内の重要な評価項目の抽出を行った研究として，複数の共起範囲に対してトピックの分布を仮定した MultiGrainLDA[Titov 2008] などが存在する．しかしトピックモデルを用いるときの問題として，トピック数を増加させた際に，類似したトピックや，意味の混在したトピックが増加してしまうことが挙げられる．

2.2 潜在的ディレクレ配分法 (LDA)

LDA は，文書内の各単語の背景に潜在変数 (トピック) を仮定し，また，文書にトピックの出現確率分布を仮定し，単語の生成過程をモデル化した代表的なトピックモデルである．LDA

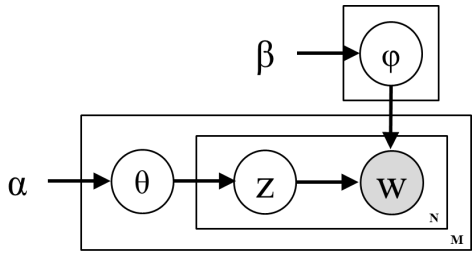

図 1: LDA のグラフィカルモデル

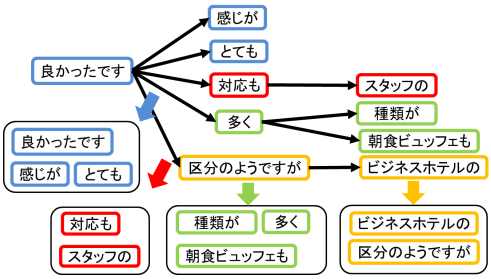

図 2: 係り受け関係

における文書の生成過程の流れを以下に示す．
(a) 文書毎に，ディレクレ分布 Dir(α) に従い，トピックの出現確率分布 Θ を生成する．
(b) トピック毎に，ディレクレ分布 Dir(β) に従い，単語出現確率分布 Φ を生成する．
(c) 文書内の単語毎に，(a) で生成したトピックの出現確率分布 Θ に従い，トピック z を生成する．
(d) 文書内の単語毎に，(b) で生成したトピックの単語出現確率分布 Φ に従い，単語 w を生成する．
上記 (c)，(d) を，全文書・全単語に関して行う．

LDA における文書の生成過程を表すグラフィカルモデルを図 1 に示す．

3. 提案手法

3.1 係り受け情報を用いたトピックの粒度の細分化

トピックモデルには，トピック数を増加させた際に，類似したトピックや，意味の混在したトピックが増加するという問題がある．これは，トピックの推定の際に，文書・文単位で単語の共起を利用していることが原因の一つであると考えられる．なぜなら，単語 w のトピックの推定は，推定される箇所以外での w に対するトピックの割り当てられ方と，トピックの分布を仮定した共起範囲における，w 以外の単語のトピックの割り当てられ方に依存するためである．そのため，文書や文に複数の評価項目に関する単語が存在する際に，それらをトピックとして細分化することが困難となっていると考えられる．ここで，文内で様々なトピックに関連する単語が混在している例文を示す．
「ビジネスホテルの区分のようですが朝食ビュッフェも種類が多く, スタッフの対応もとてもよかったです.」
文書・文の単位の集合では,「対応」と「スタッフ」,「朝食ビュッフェ」と「種類」といった，異なる評価項目に関する単語が混在し，トピックの粒度の細分化が難しくなっていると考えられる．

そこで本稿では，より局所的な単語の共起関係を考慮するために，係り受け情報を用いる．上記の文の係り受け関係を図 2 に示す．図 2 のように，係り受け関係により，単語間の修飾関係がわかる．係り受け情報を用いて文をいくつかのクラスタに分割し，局所的な単語の共起関係を考慮して，トピックの分布を仮定する．文の分割の様子を図 2 に示す．係り受け解析から得られた構文木の枝の分岐先では，単語間の修飾関係が存在する．本稿では，係り受け木の枝先に 2 単語以上存在する場合は，その枝先を “係り受けクラスタ” として文から分割し，一つのクラスタとする．また，2 単語未満の場合は，枝分かれ元に集約する．この係り受けクラスタを共起範囲とし，トピックモデルを適用する．具体的には，係り受けクラスタ毎にトピックの分布を仮定し，潜在的ディレクレ配分法 (LDA) によりトピックの推定を行う．

3.2 トピックの階層構造の構築

3.1 の提案手法を用いて，トピック数を増加させると，レビュー内容の概略が詳しく捉えられる一方で，多くのトピックが意味のまとまりを考慮されずに出力されるため，使用するユーザにとって負担となることが予想される．そこで，出力されたトピックを，大きな枠組みで類似したトピックとしてクラスタリングし，階層構造を構築する．トピックの階層関係の対応付けには cos 類似度 [Kim 2009] を用いる．ベクトル $\vec{p}, \vec{q}$ が存在するとき，それらの cos 類似度は (1) 式で表される．

$$cos(\vec{p}, \vec{q}) = \left(\frac{\vec{p} \cdot \vec{q}}{|p||q|}\right) \quad (1)$$

階層構造の構築のために，トピック数を小さくしたとき（以下，粒度の大きなトピックとする）と，トピック数を大きくしたとき（以下，粒度の小さなトピックとする）でそれぞれトピックを推定し，粒度の大きな各トピックと粒度の小さな各トピック間の cos 類似度を算出する．算出した類似度に基づき，各粒度の小さなトピックを，最も cos 類似度の高い粒度の大きなトピックに対応付けを行う．これにより，トピックの階層関係を得ることができる．

3.3 レビューの評価項目の可視化

評価項目の可視化は，初めに粒度の大きいトピックについて，単語出現確率の上位 N 個を上方に表示する．次に，ユーザにより選択された大きな粒度のトピックと，階層関係を持つ小さな粒度のトピックの単語出現確率の上位 N 個を下方に表示する．

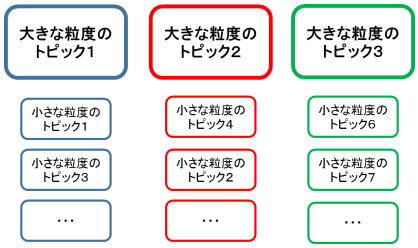

図 3: 可視化の概略図

4. 実験

4.1 係り受け情報を用いたトピック粒度細分化手法

4.1.1 評価指標

係り受け情報を用いた際のトピックの粒度の細分化について検討するために，トピックの独立性の指標として cos 類似度，均一性の指標として Coherence[Mimno 2011] を用いた．トピック数を増加させた際，各トピックが独立しており，かつ意味的に均一であれば，トピックの粒度を適切に細分化できていると考えられる．cos 類似度は，トピックの類似性の指標であり，(1) 式で表される．この指標は，分布の類似性が高いほど大きな値をとり，独立性が高いほど，小さな値をとる指標である．また，トピック t の単語の出現確率分布の m 番目に出現しやすい単語を v_m^t，単語集合内で v_m^t が出現した回数を $D(v_m^t)$，単語集合内で v_m^t,v_l^t が共起した回数を $D(v_m^t,v_l^t)$ すると，Coherence は次の式で表される

$$Coherence = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log\left(\frac{D(v_m^t, v_l^t)+1}{D(v_l^t)}\right) \qquad (2)$$

ただし，[Mimno 2011] における Coherence の定義では，D は文書内の共起でカウントを行っているが，本稿ではトピックの粒度の細分化が目的であるため，D は係り受けクラスタ単位でカウントを行う．この指標は，トピックの上位語 M が，単語集合内で頻繁に共起するほど，大きな値をとる指標である．

4.1.2 実験方法

Trip.advisor.com におけるホテル A についてのレビュー (文書数:683，文数:4,896，係り受けクラスタ数:6,244，語彙数:4,280 語，全単語数:18,274 語) を用いて実験を行った．形態素解析には Mecab，係り受け解析には Cabocha を用いた．使用した品詞は名詞のみであり，名詞が連続するものは複合語として名詞を連結し，一つの名詞とした．またトピックの分布を仮定する共起範囲が文書，文，係り受けのクラスタのそれぞれ場合で，LDA によりトピックを推定し，4.1 で述べたトピックの独立性と均一性の比較を行った．これら 2 つの指標の 100 試行の平均値をそれぞれ算出し，比較を行った．トピックの推定は GibssSampling[Griffiths 2004] を用いた．また，推定の際のパラメータは，α=0.1，β=0.1，サンプリング回数 1000 回，トピック数 10〜100(10 刻み) として実験を行った．

4.1.3 結果および考察

初めに，3.1 で述べた，文書・文内に複数の評価項目に関する単語が混在する際に，トピック数を増加させると類似したトピックが増加するという仮説を検証した．文書，文，係り受けのクラスタに対して，トピック数を 10〜100 とし，LDA によりトピックを推定した際の cos 類似度の結果を図 4 に示す．図 4 より，係り受けクラスタを共起範囲とすることで，文書や文を共起範囲とするよりも全体的に各トピック間の独立性が高くなることが確認できる．また，図 4 とトピックの目視評価により，今回の使用データにおける最適トピック数は，おおよそ 20〜30 であると思われる．

次に，Coherence の結果を図 5 に示す．上述した，最適トピック数と考えられる 20〜30 の間において，係り受けクラスタを共起範囲としたものは，文書および文よりも Coherence の値が大きく，各トピックの均一性が高くなっている．これらから，係り受け情報を用いた提案手法は，従来の手法と比較してトピックの粒度の細分化を行うことができたと考えられる．トピック数が大きいときに，文書を共起範囲として LDA を適用した場合に Coherence の高い理由は，cos 類似度の値と各トピックの定性評価などから，単語出現確率の上位語の意味はまとまっているが，類似したトピックを多く生成しているためであると考えられる．一方，トピック数が大きいときに係り受けクラスタの均一性が他よりも低いのは，係り受けクラスタ内の単語の数が他よりも少ないことが影響していると考えられる．

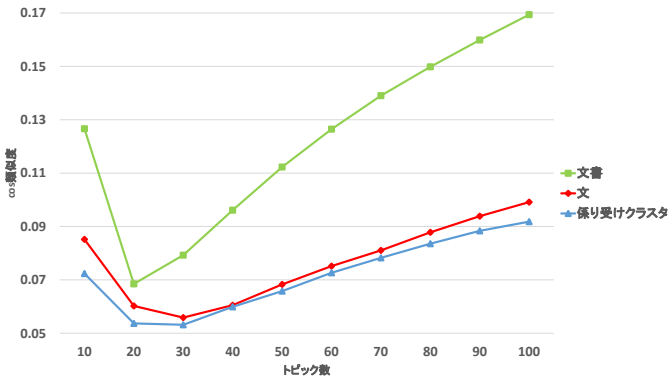

図 4: cos 類似度（独立性）

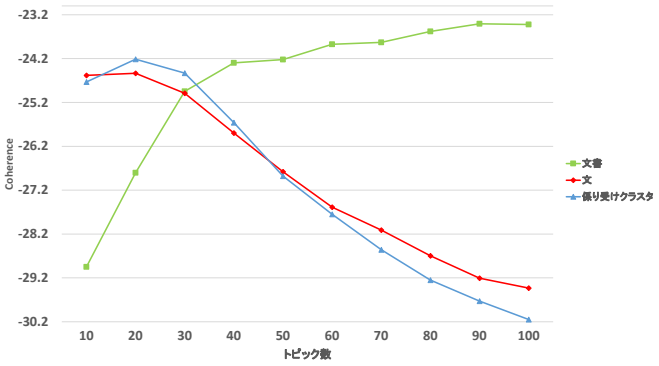

図 5: Coherence（均一性）

4.2 レビューの評価項目の可視化実験

4.2.1 実験方法

レビューの評価項目の可視化に関して検討を行うために，4.1.2 で示したレビューを用いて実験を行った．初めに，3.1 で示した方法に基づき，係り受けのクラスタを作成する．次に作成した係り受けクラスタを共起範囲として LDA を適用し，大きな粒度と小さな粒度のトピックを推定する．LDA でのトピックの推定は GibssSampling 用いて行った．

また，推定の際のパラメータは α=0.1，β=0.1，サンプリング回数 1000 回，大きな粒度のトピック数 6 個，小さな粒度のトピック数 20 個とした．それぞれのトピックの推定の後に，大きな粒度のトピックと小さな粒度のトピック間で cos 類似度を算出し，3.2 の方法に基づき，小さな粒度のトピックと大きな粒度のトピック間で階層関係の対応付けを行い，各トピックの出現確率が上位 10 個の単語を表示した．可視化結果を図 6 に示す．

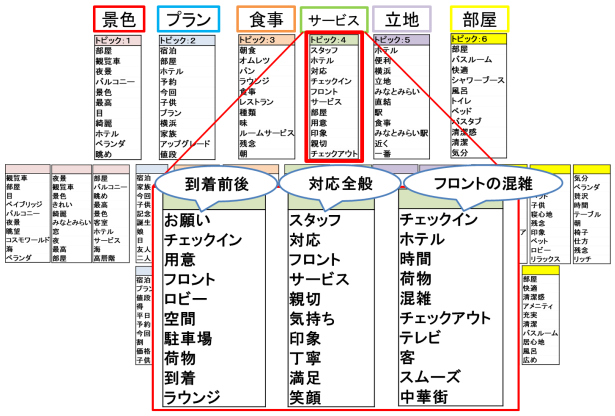

図 6: 可視化結果

4.2.2 結果および考察

図 6 より，接客全般のトピックと，より粒度の細かいチェックイン前後やスタッフの対応全般，フロントでの混雑のトピックなどが階層関係になっていることが確認できる．これは，他のトピックにおいても同様の傾向であった．この結果より，提案手法を用いることで，評価項目を階層的に表示できることが確認できた．

また，トピック数 6 のときに推定されたトピックの内容（立地，客室，サービス，食事，予約・価格，景色）と，レビューサイトにおいて付与されている評価項目（立地，客室，サービス，寝心地，価格，清潔感）が異なっている．この要因として，予想される評価項目と，宿泊者の印象に残る評価項目に差異が存在することが考えられる．

5. おわりに

本稿では，係り受け情報を用いたトピックの粒度の細分化のための手法を提案した．実験により，係り受け情報を用いることで，文書・文を共起範囲として LDA によりトピックを推定するより，各トピックの意味のまとまりと独立性が向上し，より粒度の細かいトピックが得られるを示した．さらに，階層構造を考慮したレビューの評価項目の可視化手法を提案し，内容の異なる数多くのトピックを，階層的にクラスタリングできることを示した．これにより，ユーザーが興味を持つ評価項目を容易に得られることが期待できる．

今後の課題としては，係り受けのクラスタ内に単語が少ない場合の，トピック数増加によるトピック推定性能の低下に対する検討が挙げられる．この課題に対しては，少数の単語に適したトピックの推定法を導入することなどが考えられる．

参考文献

[Blei 2003] DM Blei, AY Ng, MI Jordan: “Latent dirichlet allocation ”the Journal of machine Learning research, 2003.

[Kobashi 2005] 小林のぞみ, 乾健太郎, 松本祐治, 立石健二, 福島俊一: “意見抽出のための評価表現の収集, 自然言語処理, Vol. 12 (2005) No. 3 pp.203-222.

[Morita 2012] 森田一, 高村大也, 奥村学: “対象，属性，評価語の相互依存関係を考慮した三つ組抽出 ”, 言語処理学会第 18 回年次大会　発表論文集 2012-3 pp.723-726.

[Titov 2008] I Titov, R McDonald : “Modeling online reviews with multi-grain topic models, ”WWW ’08 Proceedings of the 17th international conference on World Wide Web, pp.111-120.

[Kim 2009] 金明哲 : “テキストデータの統計科学入門 ”岩波書店,P161,2009.

[Mimno 2011] D Mimno, HM Wallach, E Talley, M Leenders, Andrew McCallum : “Optimizing semantic coherence in topic models, ”EMNLP ’11 Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.262-272.

[Griffiths 2004] TL Griffiths, M Steyvers : “Finding scientific topics, ”Proceedings of the National Academy of sciences of the United States of America,2004.