2K-01

商品レビューの評判要因分析のための トピックモデルの適用に関する検討

月岡　晋吾 †　　吉川　大弘 †　　古橋　武 †
名古屋大学 †

1　はじめに

近年，インターネットの普及により電子商取引が増加しており，企業が自社商品と他社商品の特徴的な評判を比較することで，商品開発に応用するマーケティング戦略が活発となっている．一方でユーザレビューの投稿が増加し，レビュー全てを読み商品の評判を知ることが困難となり，テキストマイニング技術を用いた評判分析が期待されている [1]．レビューのような大量にある非構造化データを集約し，価値のある情報を抽出するデータマイニング手法の一つに Latent Dirichlet Allocation (LDA) がある [2]．またこの手法を発展させ，評判分析に適用した研究が報告されている [3]．さらに，このテキスト情報に加え，レビューの評点情報をもとに，評価に影響する評価項目（評価要因）の順位付けを行う研究も報告されている [4]．これらの手法は，基本的にはカテゴリ全体，あるいは商品単体に対する評判分析であり，商品間の比較分析は想定していない．その一方で，商品間の比較分析，特にその商品の評価に影響を与える評価要因の分析に対する重要性も指摘され始めている [5]．

本稿では，商品の機能など，評判の要因となるものを“ 評判要因”と定義し，LDA を応用した評判分析において，各商品レビューの平均評点を用いることで，複数商品の評価順位に影響する評判要因の推定を目指す．

2　評判要因分析

2.1　Latent Dirichlet Allocation (LDA)

LDA とは，レビューのような大量にある非構造化データを集約し，価値のある情報を抽出するデータマイニング手法の一つである．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することで，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 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．
(e)(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．

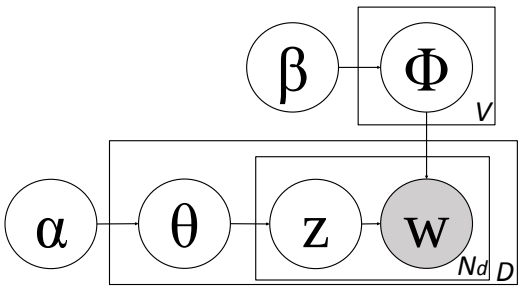

図 1: LDA のグラフィカルモデル

2.2　提案手法

ユーザによりレビューに付与された評点から商品毎の平均評点を算出する．各商品はレビューの平均評点を持ち，複数商品の評価順位に影響する評判要因の推定に利用する．次に，レビュー全体に LDA を適用し，トピックを推定することで割り当てられた各単語のトピックをカウントし，商品毎のトピック割合を算出する．これらにより得られた，各商品が持つ平均評点とトピック割合をそれぞれ標準化して，解析を行う．平均評点を目的変数，トピック割合を説明変数として重回帰分析を行うことで，評価順位に対してのトピックの影響度が偏回帰係数として得られることが期待できる．

ここで，商品毎のトピック割合の総和が 1 となるため，全てのトピック割合を説明変数として重回帰分析を行うと，多重共線性の問題が生じる．そこで，多重共線性を解消するために，一つのトピックを説明変数から削除する．本稿では，以下の手順で重回帰分析を行う．

A Study on Application of Topic Model for-Evaluation Analysis of Products

Shingo Tsukioka†，Yoshikawa Tomohiro†，Takeshi Furuhashi†，†Nagoya University

(1) トピック一つを削除し，重回帰分析を行う．
(2) 偏回帰係数の p 値>0.05 で，p 値が最大となるトピックを削除する．
(3)2 回目の重回帰分析では，(1) で削除したトピックを説明変数に追加し，重回帰分析を行う．2 回目以降の重回帰分析では，(2) 後に残っているトピックを説明変数として，重回帰分析を行う．
(4) 全ての偏回帰係数の p 値<0.05 であれば終了する．p 値>0.05 の係数が存在する場合は (2) に戻る．

3 実験

3.1 使用データ

実験には，楽天が公開している楽天トラベル [6] のホテルに関するレビューデータを用いた．宿泊目的がビジネス，宿泊人数が一人であり，東京・名古屋・大阪の中心地域のレビュー数が 100 件以上のホテルのレビュー (全 30,971 件) を対象とした．

3.2 方法

形態素解析には Cabocha[7] を用いた．名詞が連続する際は複合語とし，名詞と複合語のみを解析に用いた．ストップワード辞書は主観により作成したもの用い，全レビューデータに対して LDA を適用しトピックを推定した．推定パラメータは，α=0.1, β=0.1, サンプリング回数 1,000 回，トピック数は 10 とした．

3.3 結果と考察

推定されたトピックの上位 10 個の単語を表 1 に示す．また，表の一行目に，主観で推定したトピック名を示す．次に，重回帰分析の結果を表 2 に示す．2.2 で示した重回帰分析を適用した結果，回帰モデルは三つのモデルに集約された．商品 (ホテル) 間で平均評点，各トピック割合を標準化したため，回帰式の切片は 0 となった．本回帰モデルに対する分散分析の結果は，p 値<0.05 となった．また，表 2 に示す重相関係数と重決定係数より，得られた回帰モデルは妥当なものであると考えられる．

表 2 の結果では，悪い評価に影響するトピック (topic1) を削除すると加点モデルとなり，良い評価に影響するトピック (topic2,6,7 など) を削除すると減点モデルとなる傾向が見られた．このことから，騒音のトピックは悪い評価のホテルの特徴であり，再訪・部屋・満足性・風呂のトピックは評価の良いホテルの特徴となっていることが考えられる．

表 1: トピックの単語出現確率上位 10 個

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	topic10
騒音	再訪	清掃	部屋	対応	満足性	風呂	不明	朝食	立地
部屋	いつ	部屋	部屋	対応	部屋	浴場	部屋	朝食	便利
音	今回	風呂	ベッド	フロント	宿泊	風呂	ホテル	パン	駅
隣	宿泊プラン	残念	快適	ホテル	お部屋	部屋	立地	満足	コンビニ
ホテル	予約	シャワー	ホテル	チェックイン	ホテル	満足	満足	部屋	近く
廊下	宿泊	掃除	アメニティ	スタッフ	今回	便利	対応	種類	ホテル
窓	お願い	水	テレビ	宿泊	満足	宿泊	価格	ホテル	立地
エレベーター	快適	ホテル	残念	丁寧	快適	駅	値段	無料	部屋
壁	プラン	トイレ	フロント	時間	予約	ホテル	宿泊	サービス	場所
問題	お世話	改善	仕事	お願い	ツイン	出張	朝食	コーヒー	大阪駅
外	部屋	清掃	お部屋	今回	風呂	朝食	サービス	食事	非常

表 2: 得られた回帰モデル

モデル	(1)での削除	重相関係数	重決定係数	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10
1	topic1	0.79	0.63		0.41	0.30	0.90	0.53	1.40	1.16	0.35	0.64	0.61
2	topic2	0.79	0.62	−0.56		−0.24	0.15				−0.15	−0.22	−0.25
3	topic3	0.78	0.61	−0.46	0.15		0.33		0.46	0.23			
2	topic4	0.79	0.62	−0.56		−0.24	0.15				−0.15	−0.22	−0.25
2	topic5	0.79	0.62	−0.56		−0.24	0.15				−0.15	−0.22	−0.25
2	topic6	0.79	0.62	−0.56		−0.24	0.15				−0.15	−0.22	−0.25
2	topic7	0.79	0.62	−0.56		−0.24	0.15				−0.15	−0.22	−0.25
3	topic8	0.78	0.61	−0.46	0.15		0.33		0.46	0.23			
3	topic9	0.78	0.61	−0.46	0.15		0.33		0.46	0.23			
3	topic10	0.78	0.61	−0.46	0.15		0.33		0.46	0.23			

4 おわりに

本稿では，LDA と重回帰分析を用いた評判分析法において，各商品の平均評点を目的変数，トピック割合を説明変数として用いることで，複数商品の評価順位に影響する評判要因の推定を行った．今後の課題として，最適なトピック数やトピックの削除法に対する検討が挙げられる．

5 謝辞

本研究では，楽天株式会社から施設レビューデータを提供していただきました．深く感謝致します．

参考文献

[1] テキストマイニング技術とその応用，UNISYS TECHNOLOGY REVIEW，第 84 号，2005 in large databases, 20th VLDB, pp.487-499, 1994
[2] DM Blei, AY Ng, MI Jordan : Latent dirichlet allocation, the Journal of machine Learning research, 2003.
[3] Chenghua Lin, Yulan He : Joint sentiment/topic model for sentiment analysis, CIKM '09 Proceedings of the 18th ACM conference on Infor-
[4] Jon D. Mcauliffe, David M. Blei, " Supervised Topic Models, "Advances in Neural Information Processing Systems 20, 2007.
[5] 川中翔，宮田章裕，東中竜一郎，星出高秀，藤村考: トピックモデルを用いた消費者場面毎の競合分析, 人工知能学会全国大会論文集，25, 1-4, 2011.
[6] 楽天トラベル：http://travel.rakuten.co.jp/
[7] Cabocha : http://taku910.github.io/cabocha/