第 2 回勉強会　レジュメ　2018 年　5 月 18 日(木)

富田優

取り上げる論文について	
タイトル	：　トピックモデルを用いた商品の評判要因分析に関する検討
著者	：　月岡晋吾、吉川　大弘、古橋武
掲載誌	：　日本知能情報ファジィ学会　講演論文集 31(0), 655-660, 2015

Abstruct

近年、商品のレビューに対する評判分析の研究が行われてきた。評価分析をすることで、ユーザーは商品の概要を把握することができるし、企業は自社製品の強みと弱みを理解することができる。LDA モデルによって表現されるトピックモデルは多くの良い結果を残してきた。この論文はトピックモデルを用いて商品の評判要因を分析し、そこから得られた特徴量を用いて多重回帰分析を行い、どの要因が評点に影響を与えているかを分析した。

１．はじめに

近年、インターネットの普及で EC における購買取引が増加している。同時に商品に対するレビューも同時に投稿されており、このデータは商品やサービスに対する評判を含んでいて、この解析には多くのニーズとメリットがある。ユーザーはレビューから詳しく商品やサービスのことを知ることができる。企業は自社商品やサービスに対する評判を知り、次への改善につなげることができる。レビューにおいてはレーティングの情報もあるが、ある商品のカテゴリーにおいてどの要素がそれらの評判に大きく影響するのかを解析するのも重要である。

２．評判分析

	項目選択方式	自由記述方式
メリット	1.　モニターの負担が軽い	1.　事前に想定できなかった評判要因を知られる
デメリット	1.　事前に項目を決める必要 2.　サンプル数が必要 3.　人的労力と金銭的費用が大	1.　モニターの負担が重い 2.　解析に労力が必要 3.　多変量解析などの統計的解析手法が使いにくい

	EC サイト上のユーザーレビュー
メリット	1. 容易に多くのデータを収集可能 2. 統計処理しやすい評点情報 3. 自由記述であるレビュー情報

３．論文概要

本論文では

①評判要因（商品の機能など、ユーザーに評価されて商品やサービスの評判に影響する要素）を定義

②トピックモデルを用いてレビュー内に出現する単語をトピックとしてクラスタリング

－多くの単語が出現するレビューを次元削減して簡略に表現

－商品ごとのトピック割合を求め評判要因を求める

③重回帰分析を用いた評判要因トピックの算出

－ユーザーレビューに付与された評点から商品の平均評点を算出

－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

－平均評点を上げる傾向にあるトピックや下げる傾向にある偏回帰係数として得る

－トピック割合の総和は１なのでそのまま適用すると多重共線性の問題が発生するため、一つのトピックを削除

④評判要因の可視化

－重回帰分析で求めたトピックごとの偏回帰係数のグラフを表示

４．実験データ

楽天トラベルの施設レビュー

ビジネス目的かつ一人で宿泊したものに限定

東京・大阪・名古屋の中心地域にあり、レビュー数が 100 以上のホテルが対象

30971 件のレビューが対象

東京 46 軒、名古屋 37 軒、大阪 46 軒

総単語数 154,477

総語彙数 34,302→品詞は名詞のみ

５．数学の用意

○カテゴリ分布

複数の離散値から１つの値をとる確率分布

EX.コンビニの一番くじを１回引く、サイコロを１回振る

$$\{1,2,3,\ \cdot\ \cdot\ \cdot, V\}\ P(V) = \Phi_v$$

○ディリクレ分布

トピックモデルの事前分布に使われる

$$Dirichlet(\Phi|\beta) = \frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\prod_{v=1}^{V}{\Phi_v}^{\beta_v-1}$$

５．１ユニグラムモデル

－ユニグラムモデルを拡張したものがトピックモデル

－文書を単語の多重集合でと捉える

－多重集合・・・重複 OK な単語の集合、BOW（Bag　Of　Words）表現と呼ぶ

－形態素解析で文章を単語に分割

－パラメータ Φ が与えられたときの文書集合 W の確率は以下の通り

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{\varphi}) &= \prod_{d=1}^{D} p(w_d|\boldsymbol{\varphi}) \\ &= \prod_{d=1}^{D}\prod_{n=1}^{N_d}(w_{dn}|\boldsymbol{\varphi}) \\ &= \prod_{d=1}^{D}\prod_{n=1}^{N_d}\varphi_{w_{dn}} \\ &= \prod_{v=1}^{V}\varphi_v^{N_v} \end{aligned}$$

あとはこのΦ_vをデータから推定する

W:文書集合
Φ：φ_vのベクトル表示
w_d:文書 d の単語集合
Φ_v：単語 v が出現する確率
N_d:文書 d に含まれる単語数
w_{dn}:文書 d の n 番目の単語
$\Phi_{w_{dn}}$:文書 d の n 番目の単語が出る確率

○モデルの前提

すべての文書の単語は同一の分布から生成される

－文書ごとに異なるトピックを持っているのじゃないか

－一つの文書は複数のトピックを持っているのじゃないか

－以上の 2 点を考慮してトピックモデルに拡張

５．２トピックモデル

パラメータ Φ が与えられたときの文書集合 W の確率は以下の通り

$$p(\boldsymbol{w}|\boldsymbol{\theta_d}, \boldsymbol{\Phi})$$

$$= \prod_{n=1}^{N_d} \prod_{k=1}^{K} p(z_{dn} = k|\theta_d)\, \boldsymbol{p}(\boldsymbol{w_{dn}}|\boldsymbol{\Phi_k})$$

$$= \prod_{n=1}^{N_d} \sum_{k=1}^{K} \theta_{dk}\, \varphi_{kw_{dn}}$$

あとはこのθ_{dk}，$\varphi_{kw_{dn}}$,K をデータから推定する

パラメータの推定方法は

①最尤推定(最尤法でパラメータを推定)

②MAP 推定(事後確率が最大になるようにパラメータを推定)

③ベイズ推定（事前分布を仮定してデータを用いてパラメータの事後分布を推定）

④変分ベイズ推定

⑤崩壊型ギブスサンプリング

６．論文の結果

各トピックの単語出現確率の上位２０単語

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	topic10
騒音	再訪	清掃	部屋の設備	対応	満足性	お風呂	全体	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
立地	ホテル	今回	コンセント	チェックアウト	シングル	今回	フロント	朝	食事
残念	禁煙	お湯	加湿器	部屋	きれい	温泉	設備	メニュー	飲食店
声	出張	臭い	立地	親切	大変	いつ	非常	バイキング	快適
朝	今後	タバコ	非常	笑顔	次回	疲れ	駅	残念	出張
宿泊	ポイント	浴槽	きれい	女性	チェックイン	朝	スタッフ	サラダ	徒歩
エアコン	満足	問題	机	荷物	子供	仕事	きれい	豊富	宿泊
ドア	シングル	髪の毛	満足	残念	朝食	東京駅	出張	立地	地下鉄
空調	次回	立地	清潔	電話	対応	時間	ビジネスホテル	充実	きれい
夜	定宿	バスタブ	綺麗	サービス	アップグレード	最高	綺麗	スープ	雨
仕方	大阪	ユニットバス	風呂	仕事	綺麗	サウナ	リーズナブル	駅	大変

○重回帰分析

$$y_j = \beta_0 + \sum_{i=1}^{9} \beta_i x_{ij}$$

y_j: ホテルの平均評点

x_{ij}:　商品ごとのトピックの割合(トピック８は除く)

β_i:各トピックの偏回帰係数

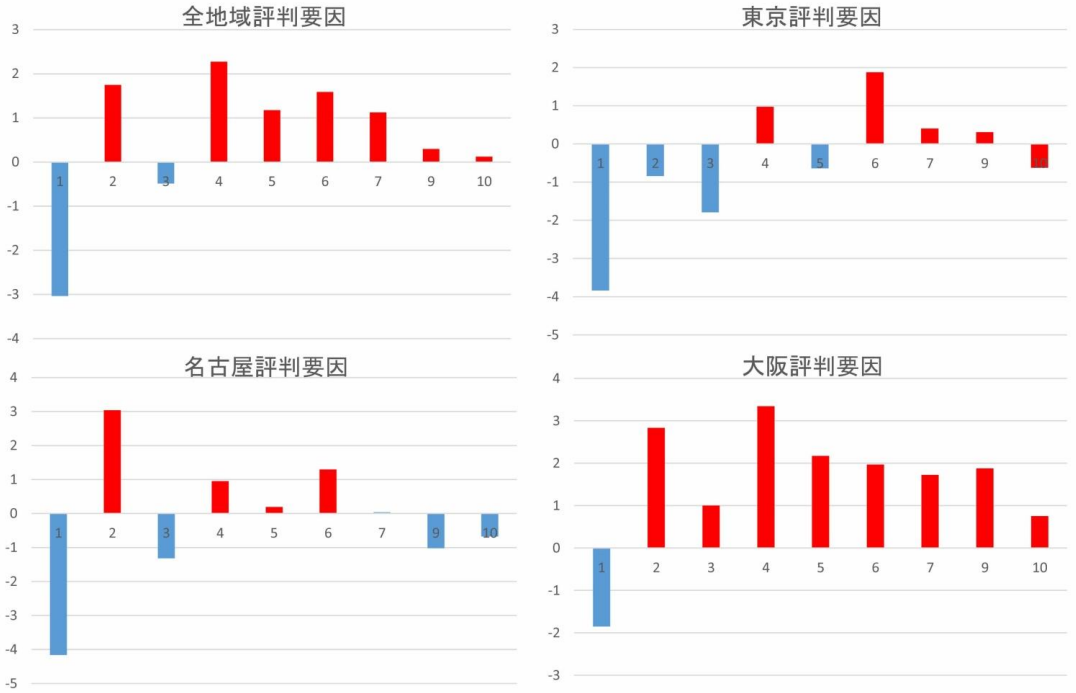
※トピックの割合すべてに回帰すると多重共線性が発生するので、今回は最も意味があいまいであるトピック８を削除

各地域の重回帰分析の結果

	全地域		東京		名古屋		大阪	
	偏回帰係数	p値	偏回帰係数	p値	偏回帰係数	p値	偏回帰係数	p値
切片	3.58	1.89E−09	4.50	3.44E−05	4.22	0.002	2.71	0.013
topic1	**−3.03**	**3.44E−04**	**−3.84**	**0.009**	**−4.17**	**0.024**	−1.85	0.151
topic2	1.75	0.081	−0.84	0.593	3.03	0.120	2.84	0.158
topic3	−0.49	0.528	−1.79	0.185	−1.32	0.464	1.00	0.483
topic4	**2.27**	**0.001**	0.97	0.387	0.95	0.570	**3.35**	**0.014**
topic5	1.18	0.138	−0.64	0.679	0.19	0.918	2.17	0.090
topic6	**1.59**	**0.006**	1.88	0.122	1.30	0.391	1.97	0.062
topic7	**1.12**	**0.047**	0.40	0.687	0.04	0.976	1.73	0.093
topic9	0.30	0.655	0.31	0.787	−1.02	0.472	1.88	0.292
topic10	0.12	0.861	−0.62	0.594	−0.68	0.690	0.75	0.578

※赤はＰ値が 0.05 未満

各地域における評判要因



名古屋における評判要因の可視化

topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic9	topic10
部屋	いつ	部屋	部屋	対応	部屋	浴場	朝食	便利
音	今回	風呂	ベッド	フロント	宿泊	風呂	パン	駅
隣	宿泊プラン	残念	快適	ホテル	お部屋	部屋	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	部屋	近く
廊下	宿泊	掃除	アメニティ	スタッフ	今回	便利	種類	ホテル
窓	お願い	水	テレビ	宿泊	満足	宿泊	ホテル	立地
エレベーター	快適	ホテル	残念	丁寧	快適	駅	無料	部屋
壁	プラン	トイレ	フロント	時間	予約	ホテル	サービス	場所
問題	お世話	改善	仕事	お願い	ツイン	出張	コーヒー	大阪駅
外	部屋	清掃	お部屋	今回	風呂	朝食	食事	非常
立地	ホテル	今回	コンセント	チェックアウト	シングル	今回	朝	食事
残念	禁煙	お湯	加湿器	部屋	きれい	温泉	メニュー	飲食店
声	出張	臭い	立地	親切	大変	いつ	バイキング	快適
朝	今後	タバコ	非常	笑顔	次回	疲れ	残念	出張
宿泊	ポイント	浴槽	きれい	女性	チェックイン	朝	サラダ	徒歩
エアコン	満足	問題	机	荷物	子供	仕事	豊富	宿泊
ドア	シングル	髪の毛	満足	残念	朝食	東京駅	立地	地下鉄
空調	次回	立地	清潔	電話	対応	時間	充実	きれい
夜	定宿	バスタブ	綺麗	サービス	アップグレード	最高	スープ	雨
仕方	大阪	ユニットバス	風呂	仕事	綺麗	サウナ	駅	大変

７．まとめ

1. トピックモデルを用いた商品の評判要因解析のための手法の提案
2. 各地域に正の評判要因と負の評判要因を確認
3. 地域ごとの比較で特徴を検討

○著者があげている今後の課題

1. 異なるサービス・商品への適用
2. 各レビュー評点に対しての重回帰分析

８．参考文献

[1]月岡晋吾、吉川　大弘、古橋武：トピックモデルを用いた商品の評判要因分析に関する検討、講演論文集 31(0), 655-660, 2015

[2]岩田具治：トピックモデル、講談社,2015