アイテム評価への影響が大きいレビューの傾向

志村 諒　三沢 翔太郎　佐藤 政寛　谷口 友紀　大熊 智子
富士ゼロックス株式会社
{shimura.ryo, misawa.shotaro, sato.masahiro, taniguchi.tomoki, ohkuma.tomoko}@fujixerox.co.jp

1 はじめに

近年の Amazon[1,2]を代表とする EC サイトではカスタマーレビュー・システムを提供している．カスタマーレビュー・システムとは，購入者からのフィードバック（**レビュー**）を，他のユーザーに共有できるシステムである．レビューは，商品の品質を表す評価値（評価値を与える行為を**アイテム評価**，その値を**アイテム評価値**と表記）とテキスト（**レビュー文**）から構成される．投稿されたレビューは，他のユーザーから投稿日時や投票数[3]と共に閲覧が可能となる．カスタマーレビューのアイテム評価値は，(1) 別のユーザーがアイテムを購入するか否かの適切な意思決定を支援し，(2) 推薦システムの高精度な推薦を可能とする．ただし，ユーザーが本来のアイテム評価値を与えていることが前提となっている．

しかし，カスタマーレビュー・システムのアイテム評価値には，過去のレビューによるバイアスが生じる可能性があることが近年の研究で示されている [1]．ここでいうバイアスとは，システムによって提供される情報に影響され，ユーザーのアイテム評価が適切に行われないことを指す．Admavicius ら [1] の研究では，同品質のアイテムを比較した時，高い平均アイテム評価値が提示されたアイテムに対して，ユーザーはより高いアイテム評価値を与えやすいことが示されている．このようなバイアスは，ユーザーの適切な意思決定を妨げ，推薦システムの高精度な推薦を困難にする．

従って，カスタマーレビュー・システムで得られるアイテム評価値からバイアスを取り除いた，真の評価値の推定が重要となる．これまでの研究 [2,3] では，平均アイテム評価値によって生じるバイアスのモデル化が取り組まれている．

では，ある時点でのアイテム評価時に，それ以前のレビューによって生じるバイアス（**レビューバイアス**）は，過去の全てのアイテム評価値がそれぞれ均等な強さで生じさせるのだろうか？ カスタマーレビューシステムでは，過去のアイテム評価値を参照する際，レビューが含むレビュー文や投稿日時・投票数も得ることができる．これらの情報を通して，過去のアイテム評価値に対するユーザーの信頼性や共感の程度などが変化し，そのアイテム評価値によるバイアスが生じると考える．例えば，商品の特徴が詳細に語られたレビューを信頼性が高いと判断し，そのレビューから強いバイアスを受ける可能性がある．また，新しいレビューの評価値は信頼性が高く，強いバイアスを生じる可能性もあると考えられる．他にも，価格を決め手にあるアイテムを購入したユーザーは，同じくレビュー文で価格について言及しているレビューに共感し，そのレビューのアイテム評価値から，強いバイアスを受けるかもしれない．このようなバイアスの強さは，何らかのレビューに関する特性（**バイアス要因**）によって決定されると考えられる（上の例では，投稿日時・レビュー文の詳細さなど）．

バイアス要因が特定できれば，従来より高い精度でバイアスのモデル化が可能となる．そこで，本研究では，以下の 3 つの特性がバイアス要因であるか検証した．

- **新しさ**：最近投稿されたレビューの信頼性が高いと考えられる．
- **詳細さ**：レビュー文でアイテムについて詳細な記述があるレビューの信頼性が高いと考えられる．
- **類似度**：ユーザーの注目している観点（機能，価格，ブランドなど）がレビュー文で言及されているレビューへ共感すると考えられる．

検証には Amazon に投稿されたレビューを用いた．そして，バイアス要因ごとにレビューバイアスの強さを定量化し，それぞれバイアス要因として作用しているか分析した．その結果，類似度はバイアス要因であることを示唆する結果を得た．

2 関連研究

アイテム評価値のバイアスの実験的研究　アイテム評価時に提示された平均アイテム評価値により，アイテム評価に偏りが生じるという結果が，被験者への応答による実験から明らかとなっている [1]．

アイテム評価値のバイアスのモデル化　アイテム評価のバイアスをモデル化する手法が提案されている．Wang ら [4] は，過去のアイテム評価値分布をもとに，次のアイテム評価値を推定する生成モデル HEARD を提案した．また，Zhang ら [2] と Liu ら [3] は，実データにおいて，ある時点でのアイテム評価値と，それ以前のアイテム評価値の平均とにある相関関係を明らかにし，過去のアイテム評価値の影響をモデル化した．同時に，過去の評価値の影響を考慮した評価値予測手法を提案し，予測性能の向上を示した．Zhang らのモデルでは，更に投稿時期による重み付け平均を用いることによる性能向上も示しているが，これらの研究は

[1] https://www.amazon.com/
[2] AMAZON は，Amazon Services LLC およびその関連会社の商標です．
[3] Amazon では，ユーザーが「役に立った」と思うレビューに対して投票を行うことができる．

バイアス要因の作用を重視したものではない．また，レビュー文も考慮されていない．

テキストを利用したアイテム評価値予測手法　レビュー文を用いた様々な評価値予測手法が提案されている [5–8]．McAuley ら [5] は LDA を用いてレビュー文のモデリングを行なっている．他の手法 [6–8] では，レビュー文を CNN や GRU でエンコードして予測モデルに取り入れることで，予測性能の向上を達成している．しかし，これらの提案手法におけるレビュー文の寄与は，ユーザー/アイテムのモデリングにある．本研究では，レビューやレビュー文から，バイアス要因となるような特性を得ることが目的であるため，これらの手法とは異なる．

3　レビューバイアスの定量分析

ある時点でのアイテム評価値と，過去の部分レビュー集合の平均アイテム評価値（**Public Opinion; PO**）との関係を回帰分析した．PO は，バイアス要因ごとに抽出した部分レビュー集合内での平均アイテム評価値とした．そして，回帰直線の傾きを比較することで，バイアス要因としての作用を検証した．

3.1　データセット

本研究では， EC サイト Amazon に投稿されたレビュー・データセット [9] を対象に分析を行った．データセットには，1996 年 5 月から 2018 年 10 月までに投稿されたカスタマーレビューを含む．本研究では，そのうち表 1 に示す 3 つのカテゴリを対象とした．

表 1. Amazon データセット概要.

	アイテム数	ユーザー数	レビュー数
Movies&TVs	182,032	3,826,085	8,765,568
Electronics	756,489	9,838,676	20,994,353
Clothing	2,681,297	12,483,678	32,292,099

3.2　部分レビュー集合の抽出

次節で算出するアイテム p の i 番目の Public Opinion $o_{p,i}$ は，$i-1$ 番目までのレビューから抽出した，同じバイアス要因を持つ部分レビュー集合 C_i 内の平均アイテム評価値である．C_i として，以下の 3 つの部分集合を抽出した：

- **新しさ**：$i-1$ 番目から遡って直近 10 件を抽出．Zhang らの研究 [2] で，最近のアイテム評価値に重み付けすることによる精度向上が確認されている．
- **詳細さ**：レビューの詳細さの指標としてレビュー文長を用い，$i-1$ 番目までに投稿されたレビューのうち，レビュー文の文字数が多い上位 10 件を抽出．
- **類似度**：$i-1$ 番目までに投稿されたレビューのうち，i 番目に投稿されたレビュー文と TF-IDF ベクトルのコサイン類似度が近いレビュー文を持つレビュー 10 件を抽出．

加えて，ベースラインとして以下の方法でも部分レビュー集合を抽出した：

- **ランダム**：$i-1$ 番目までに投稿されたレビューの中から無作為に 10 件抽出．Zhang ら [2] はそれまでの平均アイテム評価値を対象としていたが，本研究ではサンプル数を等しくするため，無作為の抽出とした．
- **有用度**：$i-1$ 番目までに投稿されたレビューのうち，投票数が多い上位 10 件を抽出．投票数は，他のユーザーからの信頼性・共感が高いシグナルであると考えた．

なお，$i \leq 11$ の場合は，$i-1$ 番目までの全てのレビューを対象とした．

3.3　回帰・相関分析

はじめに，同等の品質のアイテム集合を得るために，アイテム評価値の全体平均が $[2.9, 3.1]$ に含まれるアイテム集合 I を抽出した．

続いて，アイテム $p \in I$ の i 番目の評価値 $r_{p,i}$ に対する PO である $o_{p,i}$ を推定した．レビューを $(r_{p,i}, t_{p,i})$ という評価値 r とレビュー文 t のペアとし，アイテム p に対する全カスタマーレビュー集合を $H_p = \{(r_{p,1}, t_{p,1}), ..., (r_{p,i}, t_{p,i}), ...\}$ とする．前節で抽出した部分レビュー集合 $C_i = \{(r_{p,j}, t_{p,j}) | j < i\}$ 内での平均アイテム評価値を

$$o_{p,i} = \frac{1}{\|C_i\|} \sum_{r \in C_i} r$$

とした．I に含まれる全アイテムの全評価値に対して同様の操作を行い，PO とそれに対するアイテム評価値のペア集合 $P = \{(o_{p,i}, r_{p,i})\}$ を得た．なお，$C_i \subset H_p$ であり，共に投稿日時によりソートされた順序つき集合とする．また，PO は小数点第二位で四捨五入し，PO が取りうる全ての値の集合を $O = \{1.0, 1.1, \ldots, 4.9, 5.0\}$ とした．

最後に，任意の $o \in O$ に対して，等しい o を持つペア $\{(o, r_1), \ldots, (o, r_{n_o})\}$ を P から抽出し，ある PO (o) とその下での平均アイテム評価値（$\bar{r}_o$）のペア $\{(1.0, \bar{r}_{1.0}), \ldots, (o, \bar{r}_o), \ldots, (5.0, \bar{r}_{5.0})\}$ を得た．ここで，n_o は等しい o を持つペアの個数とし，$\bar{r}_o$ は下式で計算した：

$$\bar{r}_o = \frac{1}{n_o} \sum_{k=1}^{n_o} r_k$$

ここまでで求めた $\bar{r}_o$ は，レビューバイアスがなければ，常にアイテム評価値の全体平均である 3.0 付近の値をとるはずである．対して，レビューバイアスがあれば，PO の変化に伴って $\bar{r}_o$ も変化することになる．

そこで，PO と $\bar{r}_o$ の関係に対して線形回帰を行い，回帰直線の傾きを算出した．また，参考として [2] にならい Pearson 相関係数を算出した．ただし，サンプル数の少なさによるノイズを除くため，線形回帰および相関係数の算出には PO が $[2.0, 4.0]$ の区間のみを対象とした[4]．

これらを，3 つのカテゴリで，それぞれ正順（投稿

[4] ここで除かれるデータは全体の約 2-20%.

表 2. PO とそれに対するアイテム評価値の傾き・相関係数（上段：日付昇順，下段：日付降順）．

	傾き					相関係数				
	ランダム	有用度	新しさ	詳細さ	類似度	ランダム	有用度	新しさ	詳細さ	類似度
Movies&TVs	0.037 (0.013)	-0.019 (0.005)	0.251 (0.254)	-0.065 (0.029)	0.718 (0.428)	0.502 (0.176)	0.207 (0.061)	0.971 (0.963)	0.523 (0.451)	0.997 (0.995)
Electronics	-0.028 (-0.042)	0.013 (-0.079)	0.156 (0.156)	-0.012 (-0.009)	0.763 (0.665)	0.422 (0.771)	0.270 (0.871)	0.974 (0.977)	0.212 (0.222)	0.997 (0.996)
Clothing	-0.093 (-0.116)	-0.054 (-0.115)	0.018 (0.019)	-0.080 (-0.090)	0.629 (0.531)	0.752 (0.874)	0.624 (0.895)	0.236 (0.342)	0.694 (0.858)	0.987 (0.983)

日の昇順（旧→新））にソートしたデータと逆順（投稿日の降順（新→旧））にソートしたデータに適用した．逆順の回帰分析では，アイテム評価値と，その時点ではユーザーが閲覧できない未来に投稿された PO との関係を捉えるということになる．レビューバイアスは時間順序を逆転した場合生じないため，そこで見られる関係は疑似相関であることがわかる．

以上のことから，正順における傾きから逆順の傾き（疑似相関）を差し引いた部分がレビューバイアスであると考えた．

4 結果と考察

4.1 傾き・相関係数

表 2 各行上段に傾きと相関係数を示す．図 1 には Movies&TVs カテゴリにおける PO とその時の平均アイテム評価値の関係と回帰直線を示す．なお，Electronics カテゴリにおいても同様の傾向であったが，Clothing カテゴリにおいては一部異なる結果であった．

- **ランダム・有用度・詳細さ**：傾きはおよそ 0 であったが，相関は見られた．順序を逆順にした分析においても，傾きは変化せずほぼ 0 であった．
- **新しさ**：Movies&TVs，Electronics カテゴリで正の傾きが見られ，相関係数も非常に高かった．逆順にしても傾きは減少しなかった．Clothing カテゴリでは，正順・逆順いずれも傾きは約 0 であった．
- **類似度**：正の傾きが見られ，相関係数も非常に高かった．順序を逆順にしたことにより，傾きは正順と比較して減少が見られた．

4.2 考察

まず，傾きがおよそ 0 であったことから，有用度・詳細さはバイアス要因となっていないと考えられる．

次に，新しさは正順で高い傾きを示したが，逆順でも同程度の傾きを得たことから，疑似相関であると考えられる．よって，これもバイアス要因ではないと考えられる．疑似相関が見られた原因は，アイテム評価値に流行のような偏りがあったためと考察できる．Clothing カテゴリのデータは，アイテム数あたりのレビュー数が少ない．そのため，レビュー間の投稿期間が長くなり，投稿日時が遠いレビューも抽出された可能性がある．結果として，偏りが生じず，疑似相関が見られなかったと考えられる．

最後に，類似度は逆順で傾きが見られたが，正順ではより高い傾きを示したことから，その差分がバイアス要因としての作用であると考察した．正順・逆順両方で見られた高い傾きは，レビュー文自体が評価値と相関する語彙を含むことに起因すると考えられる．例えば，“good”/“bad” などのポジティブ/ネガティブな意味を持つ単語を含むレビュー評価値は総じて高い/低いことが予想できる．実際，レビュー文からアイテム評価値を予測する研究で，レビュー文が含む単語の直感的な意味と，アイテム評価値が相関することが示されている [10]．また，Clothing カテゴリでは，アイテム数あたりのレビュー数が少なく，類似したレビューが見つかり難かったため傾きが低かったと考えられる．

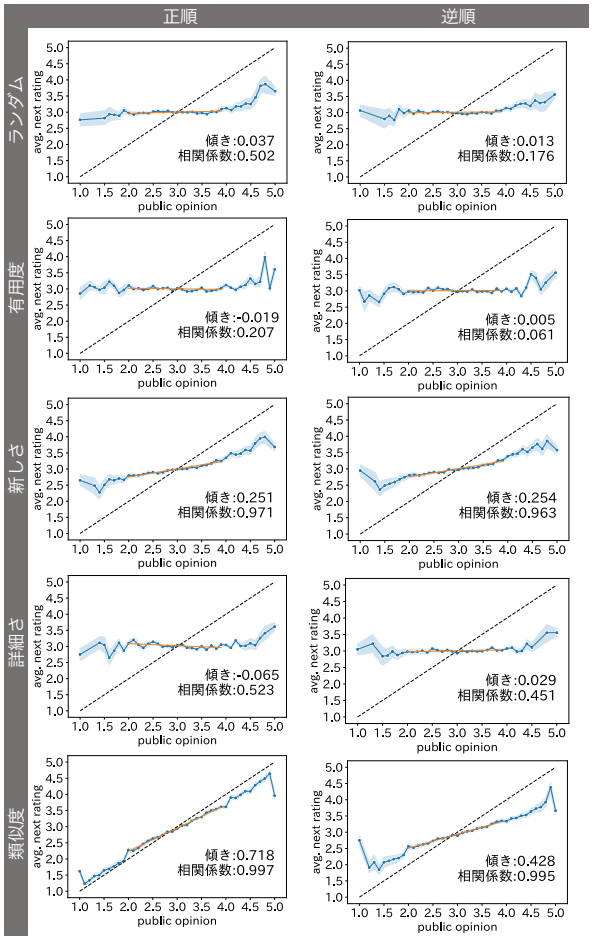

図 1. Movies&TVs カテゴリの PO とその下でのアイテム評価値の関係 (青)，回帰直線（橙）．回帰直線はサンプル数の少ない 2.0 未満，4.0 超は除いた．

4.3 類似度で抽出されたレビュー文の例示

表 3 に，レビュー文の記述内容とアイテム評価値に相関が見られる例を示す．予測対象のレビュー文で，“great” いう高評価を示す単語を含み，類似度で抽出したレビューも “great” いう単語を含む高評価値のレビューが多い（PO は 4.6）．これが，考察で述べた疑似相関の原因であると考えられる．

次に，表 4 に同じ観点のレビューが抽出されている例を示す．予測対象のレビュー文から，このユーザーは，“headband” に注目したレビューを投稿をしていることがわかる．TF-IDF の類似度によって抽出されたレビュー文は，同様に “headband” について言及しており，その PO は 2.2 である．この時点での全体平均は 3.0 であり，実際のアイテム評価値 2.0 は PO に近い値であることがわかった．この解釈として，類似度の高いレビューについて共感して，PO に同化した結果であると考察できる．

表 3. 記述内容とアイテム評価値に相関が見られる例．Electronics カテゴリのイヤホン（asin:B00006JPRQ）のレビューから抽出．レビュー文は単語に分割し，TF-IDF 降順で上位のみ掲載．★：予測対象のレビュー文．

レビュー文	評価値	類似度
★ “price, **great**”	5	-
“comfortable, sound, **great**”	5	0.52
“work, price, well”	4	0.51
“ten, son, gift, thanks, **great**, love, bought”	5	0.41
“walks, books, audio, right, price, **great**”	4	0.29
“product, **great**”	5	0.29
“work, **great**”	5	0.27
“fantastic, price, lot, quality, get, product”	5	0.27
“comfy, wear, price, good”	5	0.25
“sound, **great**”	5	0.25
“headphones/ear, pair, applies, fitted, treat, price, fitting”	3	0.24
Public Opinion/投稿時の全体平均	4.6/2.8	

5 おわりに

本研究では，Amazon のカスタマーレビュー・システムでアイテム評価時に受けるレビューバイアスについて，バイアス要因の特定を試みた．その結果，アイテム評価者のレビュー文と，それ以前に投稿されたレビュー文の類似度の高さが，バイアス要因であることを示唆する結果を得た．

今後は，特定したバイアス要因を用いてレビューバイアスをモデル化することで，アイテム評価値予測の精度向上に取り組む予定である．

表 4. 同じ観点のレビューが抽出された例．Electronics カテゴリのイヤホン（asin:B00006JPRQ）のレビューから抽出．レビュー文は単語に分割し，TF-IDF 降順で上位のみ掲載．★：予測対象のレビュー文．

レビュー文	評価値	類似度
★ “**headband**, causes, static, buds, inexpensive, metal, comfortable”	2	-
“**headband**, ridged, majority, workout, pieces, flimsy, holding”	2	0.26
“ear, buds, maxells, sound, better, hold, **headband**”	1	0.26
“metal, halo, is–or, headphones, conceptualized, **headband**, head–”	1	0.25
“hollow, decently, headphones, style, build, ok, cheap”	1	0.22
“tension, **headband**, pieces, ear, pro:, con:, enough”	1	0.22
“asked, buds, held, value, couldnt, 6, ear”	5	0.21
“wear, head, inevitably, running, securing, headphones, **headband**”	3	0.20
“maxell, buds, ears, model, -5, head, redesigned”	1	0.20
“jamming, plays, buds, ears, tell, decent, comfortable”	5	0.19
“buds, foolish, inward, aspect, pressure, comment, offers”	2	0.18
Public Opinion/投稿時の全体平均	2.2/3.0	

参考文献

[1] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. Understanding Effects of Personalized vs. Aggregate Ratings on User Preferences. In *IntRS Workshop@RecSys*, 2016.

[2] Xiaoying Zhang, Hong Xie, Junzhou Zhao, and John C. S. Lui. Understanding Assimilation-contrast Effects in Online Rating Systems: Modelling, Debiasing, and Applications. *ACM Trans. Inf. Syst.*, 2019.

[3] Yiming Liu, Xuezhi Cao, and Yong Yu. Are You Influenced by Others When Rating?: Improve Rating Prediction by Conformity Modeling. In *RecSys*, 2016.

[4] Ting Wang, Dashun Wang, and Fei Wang. Quantifying Herding Effects in Crowd Wisdom. In *KDD*, 2014.

[5] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*, 2013.

[6] Lei Zheng, Vahid Noroozi, and Philip S. Yu. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In *WSDM*, 2017.

[7] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. Neural Attentional Rating Regression with Review-level Explanations. In *WWW*, 2018.

[8] Hongtao Liu, Fangzhao Wu, Wenjun Wang, Xianchen Wang, Pengfei Jiao, Chuhan Wu, and Xing Xie. NRPA: Neural Recommendation with Personalized Attention. In *SIGIR*, 2019.

[9] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *EMNLP-IJCNLP*, 2019.

[10] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. In *COLING*, 2010.