

ユーザレビューの累積自己情報量と注文数の比較

山崎 綾一郎 † 森崎 耀太 † 照屋 良直 † 高柳 佑真 † 山岸 祐己 † ‡ §
和泉 舞 § 高林 貴仁 §

† 静岡理科大学 情報学部 ‡ 理化学研究所 革新知能統合研究センター § 株式会社良品計画 IT サービス部

1 はじめに

オンラインレビューサイトにおけるアイテムやカテゴリの評価を時系列データとして指標化することを目的とし、ユーザごとの点数確率分布を利用したレビュー点数の変換を行う。提案手法は、ユーザのレビュー投稿履歴から推定された点数確率分布における自己情報量 [1] を利用し、投稿されたレビュー点数を客観的に指標化する。この評価モデルは、回顧的 (retrospective) な枠組みによる時系列データ分析 [2] に類似した考え方で、点数確率分布の時系列変化も考慮しており、評価値を時系列データとして生成することも可能である。実験では、レビューデータセットから生成した提案評価値とレビュー点数の移動平均、およびデータセットと同時期のネットストア注文数で比較を行い、提案評価値の有用性を検証する。なお、このレビュー点数変換はエッジコンピューティング [3] による実装を目指しているため、エッジデバイスにおける高速処理と、結果の一貫性が保証されるように設計する。

2 提案手法

レビューデータセットを

$$\mathcal{D} = \{(c_1, s_{u_1,1}, t_1), \dots, (c_N, s_{u_N,N}, t_N)\} \quad (1)$$

とする。 c_n と u_n と $s_{k,n}$ と t_n は、 n 番目に投稿されたレビューの、アイテムもしくはカテゴリ $c_n \in \{1, \dots, i, \dots, I\}$ と、ユーザ $u_n \in \{1, \dots, k, \dots, K\}$ と、レビュー点数 $s_{k,n} \in \{1, \dots, j, \dots, J\} = \mathcal{J}$ と、観測時刻 $t_1 \leq \dots \leq t_n \leq \dots \leq t_N$ をそれぞれ表す。 n はタイムステップとし、 $N = \{1, 2, \dots, N\}$ をタイムステップ集合とする。便宜上、 $s_{k,n}$ は J -次元ベクトルダミー変数として

$$s_{k,n,j} = \begin{cases} 1 & \text{if } s_{k,n} = j; \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

のように変換する。このとき、ユーザ k によって m 番目のレビューで点数 j が投稿される確率 $p_{k,m,j}$ が多項

分布に従っていると仮定すると、 $p_{k,m,j}$ の最尤推定量は

$$\hat{p}_{k,m,j} = \frac{\sum_{\{n \in N \mid n \leq m\}} s_{k,n,j}}{\sum_{\{n \in N \mid n \leq m\}} \sum_{a \in \mathcal{J}} s_{k,n,a}} \quad (3)$$

のように与えられる。さらに、最尤推定量を用いたユーザ k の投稿点数確率分布を

$$\hat{\theta}_{k,m} = \{\hat{p}_{k,m,1}, \dots, \hat{p}_{k,m,j}, \dots, \hat{p}_{k,m,J}\} \quad (4)$$

とすれば、その累積分布関数

$$F(v; \hat{\theta}_{k,m}) = \sum_{\{j \in \mathcal{J} \mid j \leq v\}} \hat{p}_{k,m,j} \quad (5)$$

を考慮することができ、あるアイテムもしくはあるカテゴリ i に対して投稿されたレビュー点数 $s_{k,n}$ を、そのときのユーザ k の累積自己情報量 $-\log F(s_{k,n}; \hat{\theta}_{k,m})$ として変換することができる。レビュー点数 1 が最低評価を意味し、 J が最高評価を意味している場合、提案情報量 $-\log F(s_{k,n}; \hat{\theta}_{k,m})$ は、 m 番目のレビューを投稿したときのユーザ k にとって、点数 $s_{k,n}$ がいかに珍しく評価が低い (ネガティブ) かを示していることとなるため、以降では式 (5) を $F_{\text{neg}}(s_{k,n}; \hat{\theta}_{k,m})$ と表す。それに対し、点数 j について逆順の累積分布関数

$$F_{\text{pos}}(v; \hat{\theta}_{k,m}) = \sum_{\{j \in \mathcal{J} \mid v \leq j\}} \hat{p}_{k,m,j} \quad (6)$$

を考えれば、同様に点数 $s_{k,n}$ がいかに珍しく評価が高い (ポジティブ) かを示していることになる。ここで、 m 番目のレビューまでに 1 回しかレビューを投稿していない、もしくは複数回のレビューにおいて同じ点数 j しか投稿していないユーザ k の提案情報量は 0 になることに注意されたい。また、提案情報量は、その加法性により、時系列データとして累積し続けることも可能であると考えられる。このとき、提案モデルは観測数 N に対して $O(N)$ で全てのアイテムもしくはカテゴリの時系列データを生成可能であり、データの追加や更新があった場合でも、 $O(N)$ でそれらを再計算することが可能である。

3 評価実験とまとめ

無印良品のネットストア*におけるレビューデータセット $N = 159458$ と、データの対象期間におけるネッ

*<https://www.muji.com/jp/ja/store>

Comparing the Cumulative Self-Information of User Reviews and the Number of Orders

†Ryoichiro YAMAZAKI †Yota MORISAKI †Yoshinao TERUYA

†Yuma TAKAYANAGI ††§Yuki YAMAGISHI §Mai IZUMI

§Takahito TAKABAYASHI

†Shizuoka Institute of Science and Technology

‡RIKEN

§Ryohin Keikaku Co., Ltd.

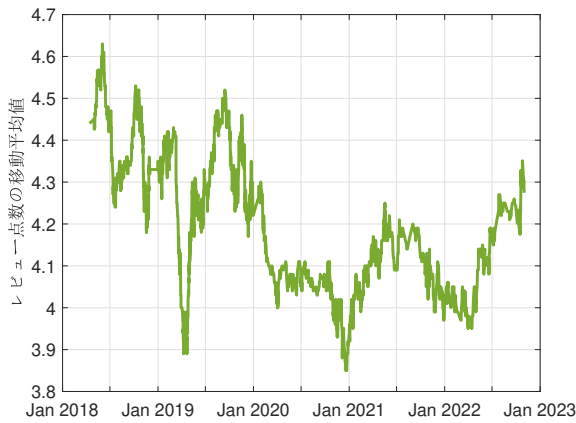
トストアの注文数を評価実験の対象とした。今回は、 w 番目までにアイテムもしくはカテゴリ i に投稿されたレビュー $N_{i,w} = \{n \in N \mid c_n = i, n \leq w\}$ におけるポジティブな意味での提案情報量の累積値

$$CP(i, w) = \sum_{m \in N_{i,w}} -\log F_{pos}(s_{k,m}; \hat{\theta}_{k,m}) \quad (7)$$

と、ネガティブな意味での提案情報量の累積値

$$CN(i, w) = \sum_{m \in N_{i,w}} -\log F_{neg}(s_{k,m}; \hat{\theta}_{k,m}) \quad (8)$$

を算出し、それらの差分 $CP(i, w) - CN(i, w)$ を提案評価値とし、レビュー点数の移動平均と比較した。以下では、提案評価値の最終値 $CP(i, N) - CN(i, N)$ が唯一正值だった 2 カテゴリ「子供服」「紳士ウェア」の結果を用いて考察を行う。



(a) 「子供服」カテゴリ

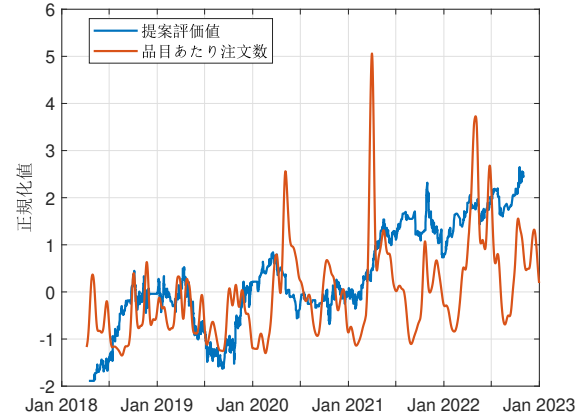


(b) 「紳士ウェア」カテゴリ

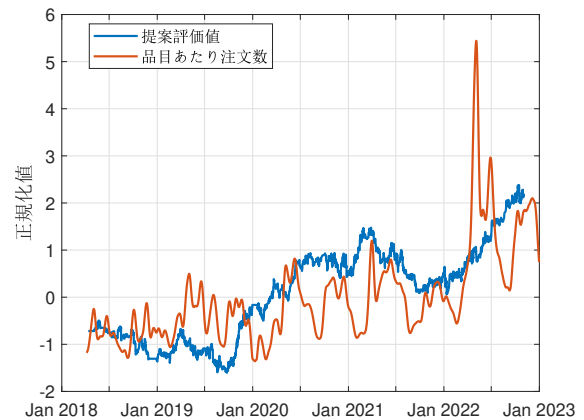
図 1: レビュー点数の移動平均値

図 1 より、両カテゴリのレビュー点数の移動平均値は、直近の回復傾向は見られるものの、全体的には減少傾向にあることがわかる。それに対し、提案評価値と品目あたり注文数の正規化値（図 2）を見ると、どちらも増加傾向にあるため、各ユーザの相対的な満足

度は高く、実際の売り上げにも繋がっていることが分かる。今後は、同様にエッジコンピューティングを想定した、標準化 z -score [4] を時系列化した値などと比較を行いたい。



(a) 「子供服」カテゴリ



(b) 「紳士ウェア」カテゴリ

図 2: 提案評価値と品目あたり注文数の比較

参考文献

- [1] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27, No. 3, pp. 379–423, 1948.
- [2] Rebecca Killick, Paul Fearnhead, and I.A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, Vol. 107, pp. 1590–1598, 12 2012.
- [3] Bin Liu, Zhongqiang Luo, Hongbo Chen, and Chengjie Li. A survey of state-of-the-art on edge computing: Theoretical models, technologies, directions, and development paths. *IEEE Access*, Vol. 10, pp. 54038–54063, 2022.
- [4] Douglas G Altman and J Martin Bland. Standard deviations and standard errors. *BMJ*, Vol. 331, No. 7521, p. 903, 2005.