

レビュー点数分布に基づく情報量の時系列データを活用した レコメンドシステムの提案

山崎 綾一郎† 能代 哲太† 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

本研究では、ユーザのレビュー点数分布に基づく情報量を利用し、レビュー点数を新たな評価指標として変換を行う。その際、ユーザの投稿点数確率分布における累積相対度数を考え、ポジティブ、ネガティブな意味での情報量の差分を算出し、時系列データとして扱う。提案指標と Diffusion Index [1] に類似したレビュー点数の累積値に対し、Dynamic Time Warping (DTW) [2] を適用し、レコメンド指標としての比較を行う。その際、ネットワーク中心性を利用し、レビュー平均点やレビュー投稿数との関連性を調査した結果、提案指標は正の相関や無相関、Diffusion Index の類似モデルでは逆相関がみられた。これにより、提案指標はメジャー商品を中心にレコメンドされるシステムの対抗策としての利用が期待される。

2 提案手法

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 [3] を利用し、投稿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いる。

また、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 \$s_{k,n,j}\$ を想定し、\$s_{k,n}\$ が点数 \$j\$ である場合には \$s_{k,n,j}\$ は 1、それ以外の場合には \$s_{k,n,j}\$ は 0 となる。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

$$\hat{p}_{k,m,j} = \frac{\sum_{\{n \in N \mid n \leq m\}} s_{k,n,j}}{\sum_{\{n \in N \mid n \leq m\}} \sum_{a \in \mathcal{J}} s_{k,n,a}} \quad (2)$$

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

$$\hat{\theta}_{k,m} = \{\hat{p}_{k,m,1}, \dots, \hat{p}_{k,m,j}, \dots, \hat{p}_{k,m,J}\} \quad (3)$$

とすれば、その累積相対度数

$$F(v; \hat{\theta}_{k,m}) = \sum_{\{j \in \mathcal{J} \mid j \leq v\}} \hat{p}_{k,m,j} \quad (4)$$

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\$ の累積相対度数の情報量 \$-\log F(s_{k,n}; \hat{\theta}_{k,m})\$ として変換できる。これを提案情報量とし、\$-\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になるため、以降では式 (4) を \$F_{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相対度数

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

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、もしくは同一の点数 \$j\$ しか投稿していないユーザ \$k\$ の提案情報量は 0 に変換される。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でき、データの更新等があった場合も \$O(N)\$ で再計算が可能である。

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 = 390993\$ 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

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

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

†Ryoichiro YAMAZAKI †Tetta NOSHIRO †‡§Yuki YAMAGISHI §Mai IZUMI §Takahito TAKABAYASHI

†Shizuoka Institute of Science and Technology

‡RIKEN

§Ryohin Keikaku Co., Ltd.

たレビュー $N_{i,w} = \{n \in \mathcal{N} \mid c_n = i, n \leq w\}$ におけるポジティブな意味での提案情報量の累積値

$$CP(i, w) = \sum_{m \in N_{i,w}} -\log F_{pos}(s_{u_m, m}; \hat{\theta}_{u_m, m}) \quad (6)$$

と、ネガティブな意味での提案情報量の累積値

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

を算出し、それらの差分 $CP(i, w) - CN(i, w)$ を提案評価値とする。ここでは、ポジティブモデルとネガティブモデルという 2 つのモデルを仮定したときの対数尤度比と似た考え方で対象を評価することを意味している。

同様に、あるアイテムもしくはあるカテゴリ i において w 番目までに投稿されたレビュー点数の累積値を

$$DI(i, w) = \sum_{m \in N_{i,w}} (s_{u_m, m} - 3) \quad (8)$$

とし、Diffusion Index の類似手法として、DI とする。

レビュー数やレビュー平均点などの単純指標との比較のため、提案評価値と DI における時系列データに対し、DTW を適用し、平均距離を算出した。次に、それぞれの平均距離を逆数としてリンクの重みを設定し、ネットワーク中心性 (PageRank, 近接中心性, 媒介中心性) を算出した。

各ネットワーク中心性指標とレビュー数およびレビュー平均点の順位関係を調査するために、Kendall の順位相関を用いた。表 1 より、PageRank (提案評価値, DI) 双方で、メジャー/マイナーが上位/下位となる傾向がないことが読み取れる。媒介中心性も PageRank と似た傾向だが、多くのアイテムでの算出結果が 0 であった。

近接中心性 (DI) と単純指標においては、マイナーアイテムが上位となる傾向が推察されるが、メジャー商品を中心にレコメンドするサイトにおいては、システムが大きく異なることが懸念される。一方、近接中心性 (提案評価値) においては、単純指標の順位構造を保ちつつも、マイナーアイテムを推薦することができ、ユーザの理解を得やすいと思われるため、DI と比べてレコメンド指標に適していると推察される。

表 1: 各中心性指標と単純指標の順位相関

指標の組み合わせ	レビュー平均	レビュー数
PageRank (提案評価値)	0.0818	-0.3597
PageRank (DI)	0.1886	-0.4272
媒介中心性 (提案評価値)	0.0122	0.0439
媒介中心性 (DI)	-0.0107	0.0499
近接中心性 (提案評価値)	0.5146	-0.0974
近接中心性 (DI)	-0.2938	-0.7552

さらに、中心性指標において特定のカテゴリが上位、下位に集中しているかを確認するために、各指標のカテゴリに対し順位和検定を行い、PageRank を例として検証を行った。

表 2 より、PageRank (提案評価値, DI) においては上位 10 カテゴリのうち 5 から 6 カテゴリが婦人カテゴリであることがわかる。これにより、婦人カテゴリがレビューの時系列傾向において代表的であることが推察される。一方、表 3 では、PageRank (提案評価値) の下位 10 カテゴリのうち 4 カテゴリが菓子であることがわかる。菓子カテゴリは、満足度に味や香りが関係し、ストアの情報からこれら要素を伝えることが難しい側面があるため、ネット上での推薦よりも実店舗で確認してもらう方が好ましいと思われる。これにより、普段見過ごしがちな他のカテゴリの商品に目を向ける機会が増え、これまで関心が持たれなかった商品への興味を促すきっかけになることが期待される。

表 2: PageRank (提案評価値, DI) 上位 10 カテゴリ

rank	category (提案評価値)	z-score	category (DI)	z-score
1	婦人カットソー	4.06	婦人ボトムス	4.77
2	婦人ボトムス	3.27	婦人シャツ	2.65
3	婦人アウタートップ	3.03	ONE-ALL	2.38
4	小物	2.55	婦人カットソー	2.27
5	収納家具	2.46	製菓材料	2.21
6	MUJIWALKER	2.45	MUJIWALKER	1.91
7	マフラー	2.41	バス・トイレ用品	1.76
8	婦人ニット	2.32	マタニティウェア	1.67
9	リビング家具	2.26	子供肌着	1.64
10	婦人シャツ	2.20	フラワー用品	1.64

表 3: PageRank (提案評価値, DI) 下位 10 カテゴリ

rank	category (提案評価値)	z-score	category (DI)	z-score
91	ケット	-1.88	筆記具	-1.79
92	チョコレート	-1.96	婦人肌着	-2.22
93	ハウスキーピング	-1.99	家電製品	-2.24
94	肌洗い小物	-2.32	タオル	-2.24
95	衛生用品	-2.39	スリッパ	-2.25
96	キッチン用品	-2.57	その他シューズ	-2.29
97	スナック	-2.67	キッチン用品	-2.31
98	レトルト	-2.76	ヘアメンテナンス	-2.54
99	ピバレッジ	-3.08	半生菓子	-3.02
100	ビスケット・クラッカー	-3.13	ハウスキーピング	-3.88

参考文献

- [1] Geoffrey H. Moore. Diffusion indexes, rates of change, and forecasting. In Geoffrey H. Moore, editor, *Business Cycle Indicators, Volume 1*, Vol. 1 of *National Bureau of Economic Research*, pp. 282–293. Princeton University Press, Princeton, 1961.
- [2] SAKOE HIROAKI and CHIBA SEIBI. Dynamic programming algorithm optimization for spoken word recognition. *Readings in Speech Recognition*, pp. 159–165, 1990.
- [3] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27, No. 3, pp. 379–423, 1948.