順序応答に基づくアイテム推薦システムの改良

IMAI, Junya / 今井, 純哉

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
60
(開始ページ / Start Page)
1
(終了ページ / End Page)
4
(発行年 / Year)
2019-03-31
(URL)
https://doi.org/10.15002/00022089

順序応答に基づくアイテム推薦システムの改良

IMPROVEMENT OF ITEM RECOMMENDER SYSTEM BASED ON ORDER RESPONSE

今井　純哉
Junya IMAI

指導教員 木村 光宏

法政大学大学院理工学研究科システム理工学専攻修士課程

In general, the item recommender systems provide the items which are preferred by the user based on some pre-collected information. Among such systems, collaborative filtering is used as one of the recommendation algorithms. The algorithm compares the preference data of the target user and other users, and estimates the favorite item to the user. In the past, the collaborative filtering has measured the user's preference by using semantic differential (SD) method. However, several recent studies showed that utilizing the order response data instead of SD method is more effective. Therefore in this study, we propose a new estimation method to improve prediction accuracy of the preference in collaborative filtering with order response information.

***Key Words** : recommender system, collaborative filtering, ranking method*

1. はじめに

(1) 研究背景

現在，情報化技術の発達に伴って人々の生活が大きく変化を続けている．その内の一つとして電子商取引が挙げられる．情報化技術と電子商取引の発展によって，個人・団体が容易かつ低コストで発信・受信が可能になった．電子商取引は店舗に行かずとも商品を購入することが可能となり，近隣に店舗がない人々等，多くの人々の利便性向上に大きな役割を果たしている．しかし，市場規模が拡大するにつれて商品取り扱い品目は増加し，電子上にある全ての商品を全て把握する事は一般的な消費者には不可能となっている．つまり，商品情報が大量に発信され，誰もが大量に情報を取得できるという情報過多に陥っている．

そこで重要となるのが推薦システム (Recommender System) である．日常生活において，大量にある情報から自身が必要とする情報に辿り着けなかったり，違いが判別不能なことは往々にしてある．そうした状況において我々は商品の口コミ，レストランのガイドブック，評論家などの他人からの推薦によって自身の行動を決定することを行っている．推薦システムは，これらの他人からの推薦情報を基に，利用者に代わって情報を探索・提供するシステムである．つまり，推薦システムは大量の商品情報から利用者が必要とするアイテムを見つけ出して提供するものである．これにより一般的な消費者が膨大な品目の中から商品を探す際の手助けとなる．

(2) 研究目的

アイテム推薦システムに用いられる手法の一つとして協調フィルタリングがある．協調フィルタリングは，他の利用者と自身の嗜好を比較して，類似した嗜好の利用者を探し出し，その利用者の好むアイテムを推薦するというシステムである．これは，自身と似た嗜好をもつ人物は別のアイテムに対しても類似した嗜好を示すという口コミの原理を用いている．協調フィルタリングにおいて嗜好を測るのに使用される手法は多くの場合，SD(Semantic Differential) 法と呼ばれる両端を対義語で表した評価尺度によって利用者の嗜好を計測しているが，その他にアイテムを好む順に並べることで嗜好を測る順位法を用いた研究も行われている．

本研究では文献 [1][2] を基に，順位法を用いた協調フィルタリングに対して新たな推定手法を用いた嗜好計算の方法を提案し，先行研究との予測精度の比較を行うことで，その有用性を検証する．

2. 順序応答に基づく協調フィルタリング

(1) 諸量の定義

本研究で用いる変数を以下のように示す．

a：利用者 a（活動利用者）
x_j：アイテム j
X_i：利用者 i の評価したアイテム集合
$O_i = x_1 \succ x_2 \succ \cdots \succ x_{|X_i|}$：利用者 i の X_i に対しての嗜好順序
$r(O_i,x_j)= r_{ij}$：利用者 i のアイテム j の嗜好順位
$D_s = \{O_1, \cdots, O_{|D_s|}\}$ ：順序 O_i の全集合

(2) データセット [3]

本研究では，寿司ネタの嗜好データを研究対象として用いる．データセットは回答者 5,000 人と 100 種類の寿司ネタに対して，10 種類のアイテムに対する嗜好順序を得る．10 種類の寿司ネタについては，寿司店での提供頻度に比例する確率で回答者毎にランダムに選ばれた寿司ネタについて評価したものである．

(3) 先行研究での手法 [2]

システムが推薦を実行するためには，推薦対象である利用者（以下活動利用者とする）と他の利用者との嗜好の類似性を測る必要がある．そのため，データベース（以下 DB とす

る）内にある他の利用者と嗜好を比較する．ここで，先行研究で最も予測精度が優れていた類似度の算出方法について明記する．嗜好の類似度は，活動利用者 a と DB 中の利用者 i とのユークリッド距離を用いて次式で測るものとする．順序 $O_i = (r_{i1}, \cdots, r_{in}), O_a = (r_{a1}, \cdots, r_{an})$ に対して，

$$D_{ai} = \sqrt{(r_{i1} - r_{a1})^2 + \cdots + (r_{in} - r_{an})^2} \tag{1}$$

これを次の類似度に変換する．

$$R_{ai} = 1 - \frac{2 \times D_{ai}}{\max_i D_{ai}} \tag{2}$$

一方の利用者の順序にしか含まれないアイテムがあった場合，そのアイテムは除外するが，その他のアイテムに対する順位の付け直しはせず，欠損した順序のまま計算を行う．

次に，活動利用者のアイテム j についての嗜好は次式で予測する．

$$P_{aj} = \tilde{r}_a + \frac{\sum_{i \in I_j} R_{ai}(\bar{r}_i - r_{ij})}{\sum_{i \in I_j} |R_{ai}|} \tag{3}$$

ただし，$\tilde{r}_a = |X_a|^{-1} \sum_{x_j \in X_a} r_{aj}$，$I_j = \{i | O_i \in D_s \text{ s.t. } x_j \in X_i\}$ である．ここで求めた P_{aj} の値が高い順にアイテムを整列したものが，システムが予測した活動利用者の嗜好順となる．

(4) 本研究での手法

本研究で提案するモデルを以下に述べる．

a) 類似度の算出

活動利用者と DB の各利用者における類似性を数値化するための計算手法として 2 変量間の距離に着目し，ミンコフスキー距離，P ミンコフスキー距離を新たな推定手法として用いた．これらの距離による類似度を以下の式で定義する．

●ミンコフスキー距離

順序 $O_i = (r_{i1}, \cdots, r_{in}), O_a = (r_{a1}, \cdots, r_{an})$ に対して，

$$D_{ai} = \left(\sum_{k=1}^{n} |r_{ik} - r_{ak}|^p \right)^{\frac{1}{p}} \tag{4}$$

ただし，p は 1.0 以上とする．ここで，式 4 を次の類似度に変換する．

$$R_{ai} = 1 - \frac{2 \times D_{ai}}{\max_i D_{ai}} \tag{5}$$

● P ミンコフスキー距離

順序 $O_i = (r_{i1}, \cdots, r_{in}), O_a = (r_{a1}, \cdots, r_{an})$ に対して，平方ユークリッド距離を用いると

$$D_{ai} = \sum_{k=1}^{n} |r_{ik} - r_{ak}|^2 \tag{6}$$

となる．ここで，式 (6) を

$$D_{ai} = \sum_{k=1}^{n} |r_{ik} - r_{ak}|^p \tag{7}$$

としたものを，P ミンコフスキー距離とする．ただし，p は 1.0 以上とする．これを次の類似度に変換する．

$$R_{ai} = 1 - \frac{2 \times D_{ai}}{\max_i D_{ai}} \tag{8}$$

b) 属性別による DB 分割

協調フィルタリングでは,「活動利用者と似た嗜好を持つ者は，活動利用者の未知アイテムに対しても類似した嗜好を持つ」という仮説に基づいている．そこで，利用者が持つ属性データによる分割を行い，活動利用者との比較対象を限定することで嗜好予測の変化を検証する．分割にはデータセットに付随する属性の内，各都道府県データを用いた．その中から，海に面している県に居住，海に面していない県に居住に分割し，検証を行った．これは，海に面していない県に居住している利用者は海鮮を食す機会が少ないが故，海鮮を好むという考えに基づいている．以下に分割に使用した各都道府県について示す．

表 1 分割に使用した各都道府県

C 県	栃木県，群馬県，埼玉県，長野県，山梨県，岐阜県，滋賀県，奈良県の海に面していない 8 県
N 県	上記以外の海に面している 39 県

ただし，C 県は 15 歳までに最も長く住んでいた県，N 県は現在住んでいる県を表している．

c) 類似度による DB 分割

活動利用者との類似度別に DB の利用者を分割し，比較対象を限定した際の予測精度の推移について検証する．提案する手法として，活動利用者との類似度が 0.2 以上，0.5 以上の利用者に DB を分割する．また，活動利用者が上位に評価しているアイテムを DB の利用者が下位に評価していた場合，利用者間の類似度がマイナスであれば，それは活動利用者に推薦すべきアイテムとなる．そのため，利用者間の類似度が-0.2 以下 0.2 以上と，-0.5 以下 0.5 以上に DB を分割した際の予測精度の推移についても検証を行う．

d) アイテム属性の追加

アイテムに対する直接的な評価データが存在する場合は予測精度は当然向上していく．しかし，アイテムに関する直接的な評価データでなくとも嗜好との相関がある情報を取得した場合も，不足した嗜好の情報量を補うことができ，予測精度の改善を見込めると考えられる．追加する情報として，寿司ネタに付随するスコアである“食べる頻度”を用いる．食べる頻度とは，ある寿司ネタをどのくらい食べるかを示す指標であり，データセットの回答者 5,000 人に対して，SD 法を用いて 100 種類全ての寿司ネタについて評価を行った平均値である．0- 3 までの数値で表されており，値が大きいほど寿司ネタが良く食べられていることになる．食べる頻度を追加した嗜好値の推定手法として，式 (3) に食べる頻度の重みを加えた次式を定義する．

$$P_{aj} + \frac{Z_{M_a} \times S_j}{\alpha} \tag{9}$$

ただし，M_a は活動利用者の“食べる頻度”であり，評価した寿司ネタの食べる頻度に対して，順位の高さを重みとした加重平均を取ったものである．Z_{M_a} は 0-1 で正規化された活動利用者の食べる頻度を表している．S_j はアイテム j の食べられる頻度を表している．α は任意のパラメータとなっており，嗜好における食べる頻度の占める割合を決定するのに用いる．

e)　未評価アイテムのみの順序予測

前節までは入力した m 個の順位を元に，入力した m 個を含む 10 種類のアイテム順序を予測していた．これは 100 種類の中からシステムがランダムに 10 種類のアイテムを提示しており，利用者が 1 位に評価したものが“本当に 1 番好き”な寿司ネタとは限らない．そのため，システムが別のアイテムを提示した際にも順序予測が行えるよう，10 種類全ての寿司ネタ順序予測を行っていた．本節では，入力順序に用いた m 個のアイテム順序を利用者が“本当に好きな順番”だと仮定し，未評価アイテムである $10-m$ 個の順序予測を行う．

(5)　実験の手順

本研究での実験の手順を以下に示す．

1. 利用者 DB から O_i を取り出し，活動利用者 a の嗜好順序 O_a とする．
2. 取り出された順序 O_a の前から m 番目，もしくは後ろから m 番目までの順序を入力とする．活動利用者 a が評価した m 個のアイテムと他の利用者 i が評価した 10 種類のアイテム順序 O_i において，2 種類以上のアイテムが重複した場合，順序 O_a と O_i の類似度を求める．ただし，m は活動利用者 a の評価したアイテム数とし，2 から 10 までの任意の整数である．
3. 求めた類似度を基にして，各アイテムの嗜好値を算出する．そして，算出された嗜好値を大きい順に整列した順序を予測順序 O'_a とする．
4. 入力に用いた活動利用者 a の嗜好順序 O_a を正確な嗜好順序とし，予測順序 O'_a を O_a と比較して発生した順序の差をシステムの予測精度と解釈し，予測精度 ρ とする．予測精度 ρ の算出には先行研究にならい，O_a と O'_a 間の Spearman の順位相関係数も用いることにした．
5. 5,000 人の利用者全ての ρ を求め，その平均を評価されたアイテム数 m における予測精度 $\bar{\rho}$ とする．

予測精度 ρ の算出については，先行研究と同様に Spearman の順位相関係数を用いて次式で測るものとする．

$$\rho = \frac{\sum_{x_j \in X_{ai}} (r_{aj} - \overline{r_a})(r_{ij} - \overline{r_i})}{\sqrt{\sum_{x_j \in X_{ai}} (r_{aj} - \overline{r_a})^2} \sqrt{\sum_{x_j \in X_{ai}} (r_{ij} - \overline{r_i})^2}} \qquad (10)$$

ただし，$X_{ai} = X_a \cap X_i$ であり，$\overline{r_i} = |X_{ai}|^{-1} \sum_{x_j \in X_{ai}} r_{ij}$ である．

3.　本研究の結果と先行研究との比較

本研究ではミンコフスキー距離，P ミンコフスキー距離を元に各手法の検証を行う．今回の実験では，p の値を 1.0 から 4.0 までの間を 0.1 刻みで変化させ，予測精度の推移について観察する．

●ミンコフスキー距離

ミンコフスキー距離による結果を図 1 に示す．図 1 は，縦軸が予測精度 $\bar{\rho}$，横軸がミンコフスキー距離のパラメータ p を表している．破線は先行研究での各 m の結果であり，実線は本手法での各 m の予測精度の推移を示している．各実線に打たれている丸点は，その m において最も予測精度が高いことを表している．

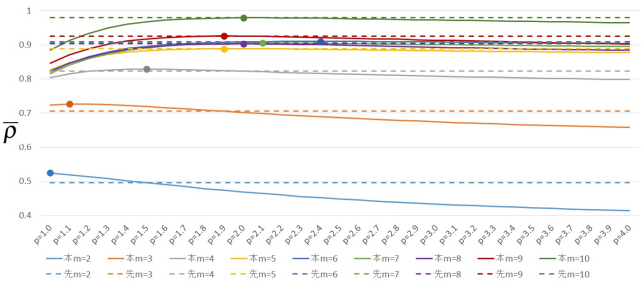

図 1 ：ミンコフスキー距離の予測精度の推移.

●P ミンコフスキー距離 P ミンコフスキー距離による結果を図 2 に示す.

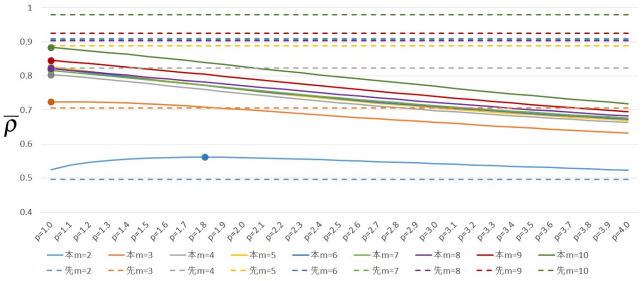

図 2 ：P ミンコフスキー距離の予測精度の推移.

●属性別による DB 分割

C 県と N 県に関係なく海あり県・海なし県に分割した，つまり DB の利用者が年齢に関係なく一度でも海のある県に居住したことがある・居住したことがないかに分割した際の結果について図 3 に示す.

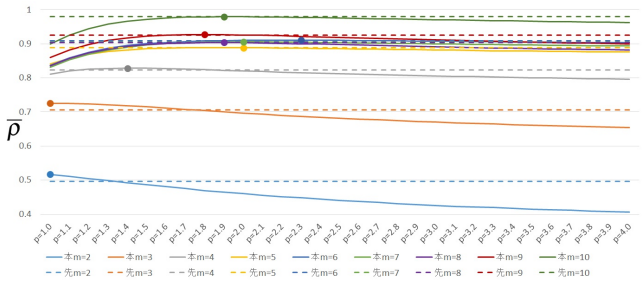

図 3 ：海ありなし CN 県に分割したミンコフスキー距離の予測精度の推移.

●類似度による DB 分割

利用者間の類似度が-0.5 以下 0.5 以上に DB を分割した際の結果について図 4 に示す.

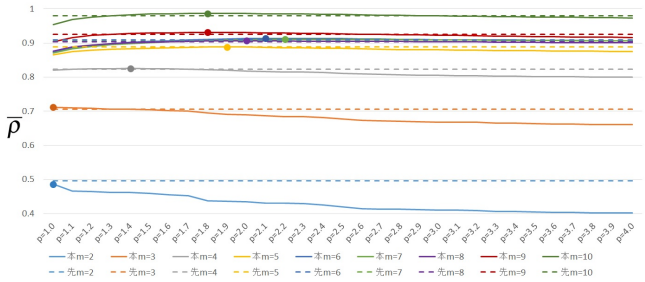

図 4 ：類似度-0.5 以下 0.5 以上に分割した分割したミンコフスキー距離の予測精度の推移.

●アイテム属性の追加

食べる頻度を加えた際の結果について図 5 に示す.

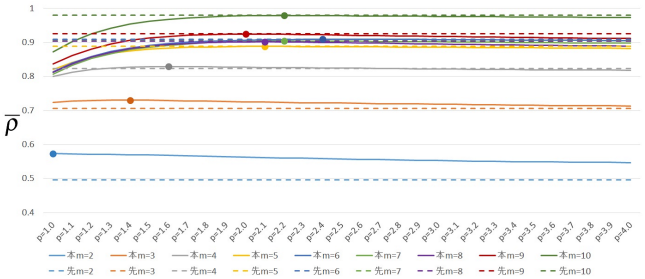

図 5 ：ミンコフスキー距離に食べる頻度を加えた予測精度の推移.

●各手法における数値結果

各手法における予測精度の数値結果を表 2，表 3 に示す．表 2 の 数字 は各 m で最も予測精度が高いことを示している．表 3 は，m 個のアイテムを入力として残りの $10 - m$ 個の順序予測を行う．$m = 9$ では 9 個のアイテムを基に残りの 1 個のアイテム順序を予測することになるが，これは必然的に 10 位となる．そのため，予測精度は 1 となる．また，$m = 10$ では予測するアイテムがないので数値は表記されていない．

1. 先行研究
2. ミンコフスキー距離
3. P ミンコフスキー距離
4. 後方からのミンコフスキー距離
5. 後方からの P ミンコフスキー距離
6. 海ありなし C 県ミンコフスキー距離
7. 海ありなし N 県ミンコフスキー距離
8. 海ありなし CN 県混在ミンコフスキー距離
9. 類似度 0.2 以上ミンコフスキー距離
10. 類似度-0.2 以下 0.2 以上ミンコフスキー距離
11. 類似度 0.5 以上ミンコフスキー距離
12. 類似度-0.5 以下 0.5 以上ミンコフスキー距離
13. 食べる頻度を加えたミンコフスキー距離

表 2 ：先行研究と本研究での予測精度の数値結果.

No.	m=2	m=3	m=4	m=5	m=6	m=7	m=8	m=9	m=10
1	0.49592	0.7053	0.82319	0.8883	0.90934	0.90473	0.90287	0.92572	0.97931
2	0.525285	0.726434	0.828933	0.888342	0.910199	0.905168	0.90287	0.926175	0.979312
3	0.561743	0.724196	0.803382	0.827028	0.823224	0.816781	0.82208	0.846051	0.879018
4	0.612036	0.736257	0.821956	0.873896	0.875004	0.84151	0.836436	0.874907	0.94497
5	0.644945	0.735993	0.792567	0.845869	0.850298	0.819353	0.791464	0.786754	0.76953
6	0.515421	0.725343	0.828478	0.888402	0.910388	0.905704	0.903455	0.927016	0.979685
7	0.515859	0.725299	0.828284	0.888516	0.910402	0.90577	0.903559	0.927207	0.979704
8	0.516686	0.725438	0.828436	0.88719	0.910395	0.905069	0.90352	0.927176	0.979653
9	0.534376	0.714601	0.81031	0.854162	0.864623	0.866897	0.876802	0.90769	0.963869
10	0.519988	0.724514	0.829949	0.889641	0.910541	0.906082	0.903741	0.92697	0.980407
11	0.484958	0.697554	0.811515	0.869181	0.891488	0.893908	0.898572	0.927067	0.985532
12	0.48609	0.712051	0.825467	0.888133	0.91312	0.91064	0.906264	0.930776	0.986344
13	0.570735	0.730119	0.829149	0.888276	0.908994	0.903668	0.901481	0.924482	0.978756

14. 未評価アイテムのみのミンコフスキー距離
15. 未評価アイテムのみに食べる頻度を加えたミンコフスキー距離

表 3 ：先行研究と本研究での未評価アイテムのみの順位予測を行った予測精度の数値結果.

No.	m=2	m=3	m=4	m=5	m=6	m=7	m=8	m=9	m=10
1	0.49592	0.7053	0.82319	0.8883	0.90934	0.90473	0.90287	0.92572	0.97931
14	0.577956	0.747571	0.842916	0.907047	0.951215	0.979488	0.994558	1	×
15	0.639084	0.755181	0.843045	0.907178	0.951154	0.979554	0.994596	1	×

4. おわりに

本研究では協調フィルタリングの予測精度改善のために，順位法を用いた推定手法によってアイテム順序予測を行った．検証の結果として，先行研究の予測精度を上回るモデルを確認することができた．特に，アイテム評価数 m が少ない場合において先行研究よりも予測精度が大きく向上したモデルもあり，大きな収穫となった．今後の課題としては，本研究では 1 つのデータセットを用いて順序予測を行ったが，他のデータセットを用いることで本研究での各モデルが適応可能なのか検証したい．

参考文献

[1] 神嶌敏弘,「なんとなく協調フィルタリング—順序応答に基づく推薦」, 2005 年度人工知能学会全国大会資料 (2004).

[2] 金高湧樹, 原尚輝,「順序応答に基づく協調フィルタリングにおける情報量の変動を考慮した推定手法の提案」, 法政大学理工学部, 平成 29 年度卒業論文 (2017).

[3] www.kamishima.net/sushi/SUSHIPreferenceDataSets (2018.08 閲覧).