

原著論文

関係性マイニングと協調フィルタリングを用いた情報推薦手法

山脇 淳一*, 工藤 康生*, 村井 哲也**

* 室蘭工業大学, ** 千歳科学技術大学

Recommendation Method Based on Interrelationship Mining and Collaborative Filtering

Junichi YAMAWAKI*, Yasuo KUDO* and Tetsuya MURAI**

* Muroran Institute of Technology, 27-1 Mizumoto-cho, Muroran-shi, Hokkaido 050-8585, Japan

** Chitose Institute of Science and Technology, 758-65 Bibi, Chitose-shi, Hokkaido 066-0012, Japan

Abstract : User-based collaborative filtering is one of the most popular recommendation methods, however, it has been pointed out that it is difficult to provide recommendation results with good recommendation accuracy and high recommendation variety simultaneously. To recommend good items for a target user, we consider that the tendency of the target user's *Kansei* evaluation to items should be explicitly reflected to recommendation methods. In this paper, we focus on pairs of items that there are big differences between target user's evaluation scores of the two items. We regard such pairs as the target user's preference patterns, and in this paper, we propose a revised user-based collaborative filtering approach that reflects the tendency of target user's *Kansei* evaluation to recommendation. The proposed method is based on explicit extraction of target user's preference patterns as interrelated attributes in rough-set-based interrelationship mining and comparison of preference patterns between the target user and other users instead of direct comparison of evaluation scores of items. Experimental results indicate that, in comparison with collaborative filtering, our proposed method can recommend appropriate items for users with at least equal or better accuracy and high variety.

Keywords : *Interrelationship mining, Collaborative filtering, Recommendation method, Preference pattern*

1. はじめに

協調フィルタリングなどの手法に基づく情報推薦技術は、現在多くの分野で注目されている。一般的に、ユーザベースの推薦手法ではユーザがアイテムに対して与えた評価値を基に、評価指標を用いてユーザ間の類似度を求めることで情報推薦を行っている。ユーザベース協調フィルタリング [1] は、ユーザベースの推薦手法の中で最も多用される手法の1つであるが、推薦の精度と多様性の両立が困難であることが指摘されており [2]、推薦精度と推薦の多様性の両立を目的とした研究が多数行われている [3-7]。ユーザベース協調フィルタリングの枠組みで推薦の精度と多様性を両立させる方策の1つとして、ユーザがアイテムに対して与えた評価値からユーザの感性的な価値判断の傾向を読み取り、情報推薦に反映させることが考えられる。例として、ユーザがアイテムに対して与えた評価値に関して、同一ユーザ内でアイテム間での評価値の差が大きい部分に、そのユーザの感性的な価値判断が強く反映されていることが予想される（例えば映画に対する5段階評価で、大好きな映画Aには評価値5を与えるが、とてもつまらないと感じた映画Bには評価値1を与えるなど）。

本研究では、このように評価値の差が大きい部分をそのユーザの「選好パターン」と呼ぶこととし、情報推薦を受ける対象ユーザと他のユーザ間で複数個の選好パターンを比較する

ことにより、ユーザベース協調フィルタリングに対象ユーザの感性的な価値判断の傾向を明示的に反映させる手法を提案する。具体的には、評価値を用いてユーザ間の類似度を直接求めることに代えて、関係性マイニング [8] の手法を用いて、2個のアイテム間での評価値の差が大きい選好パターンを明示的に選定し、ユーザ間の選好パターンの類似度を用いることで、ユーザベース協調フィルタリングを改良した推薦手法を提案する。なお、本論文は著者らの口頭発表原稿 [9, 10] を統合・拡張したものである。

2. 協調フィルタリング

本節では協調フィルタリングについて概略を述べる。本節の内容は文献 [11] に基づく。

2.1 評価値行列

協調フィルタリングの基本的な考え方は、既存のユーザコミュニティの過去の振る舞いや意見を用いて、システムを利用する対象ユーザの好みや興味のある情報を予測することである [11]。そのために、入力としてユーザの各アイテムに対する評価値を格納した評価値行列 R を用いる。ユーザの集合 $U = \{u_1, \dots, u_n\}$ とアイテムの集合 $P = \{p_1, \dots, p_m\}$ に対して、評価値行列 R は $n \times m$ 行列となり、その i 行 j 列目の値 $r_{i,j} \in \text{Range} \cup \{*\}$ は i 番目のユーザによる j 番目のアイテムの評価値を表す。ここで、 Range は評価値として取りうる値の

表1 評価値行列の例

ユーザ	アイテム					
	A	B	C	D	E	F
ユーザ1	3	5	1	*	5	*
ユーザ2	3	4	*	1	5	5
ユーザ3	4	*	2	4	4	3
ユーザ4	2	4	2	5	4	3
ユーザ5	*	4	4	3	*	1

集合を、記号*はユーザがアイテムを未評価であることを表す。評価値行列の例を表1に示す。

2.2 ユーザベースの最近傍推薦

ユーザベースの協調フィルタリングでは、対象ユーザの過去の嗜好と類似するユーザ（最近傍ユーザ）を抽出し、対象ユーザが未評価のアイテムに対する評価値を、最近傍ユーザの評価値とユーザ間の類似度を用いて予測する。ユーザ間の類似度を測るために様々な評価指標が考案されているが、代表的な指標にピアソン相関係数がある。評価値行列 R におけるユーザ a とユーザ b のピアソン相関係数による類似度は以下の式(1)で定義される。

$$\text{sim}(a,b) = \frac{\sum_{p \in P_{a,b}} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P_{a,b}} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P_{a,b}} (r_{b,p} - \bar{r}_b)^2}} \quad (1)$$

ここで、 $P_{a,b}$ はユーザ a とユーザ b が共に評価を与えているアイテムの集合である。また、 \bar{r}_a はユーザ a の評価値平均を表す。これを対象ユーザと他のユーザそれぞれについて求め、最近傍ユーザの集合 N を求める。

N を用いて、対象ユーザ a の未評価アイテム p に対する評価値を予測する。予測値 $\text{pred}(a,p)$ は以下の式(2)で定義される。

$$\text{pred}(a,p) = \bar{r}_a + \frac{\sum_{b \in N} \text{sim}(a,b) \times (r_{b,p} - \bar{r}_b)}{\sum_{b \in N} \text{sim}(a,b)} \quad (2)$$

以上の手順を用いて対象ユーザの未評価アイテムそれぞれに対して予測を行い、高い予測値を持つアイテムを推薦リストに加えることができる。

2.3 最近傍ユーザの選択

一般的に、最近傍ユーザとして対象ユーザ以外のすべてのユーザを考慮すると、対象ユーザとの類似性が低いユーザも含まれるため、推薦の質に悪影響を及ぼす恐れがある。そのため、対象ユーザと高い類似度を持つユーザのみを最近傍ユーザとして選択する必要がある。

最近傍ユーザの選択には、類似度の閾値 t を設け、 $\text{sim}(a,b) \geq t$ となるユーザ b のみを最近傍に含める、 $\text{sim}(a,b)$ を降順に並べた上位 k 人を最近傍とする、などの方法がある。どのような手法を選択するにしても、最近傍のサイズが大きすぎると予測にノイズが混じる可能性が、最近傍のサイズが小さすぎると十分に類似しているユーザを除いてしまう可能性が考えられる。

3. 関係性マイニング

本節では、関係性マイニングについて概略を述べる。本節の内容は文献[12]に基づく。

3.1 情報表と決定表

関係性マイニングは、Pawlak[13, 14]によって提唱されたラフ集合論に基づいた手法である。ラフ集合論では、データを情報表、あるいは決定表という形で用いる。情報表はYao et al.[15]によるより一般的な表記法を用いて、以下の式(3)で定義される。

$$S = (U, AT, \{V_a | a \in AT\}, R_{AT}, \rho) \quad (3)$$

ここで、 U は対象の有限集合、 AT は属性の有限集合、 V_a は属性 a の値の集合である。 $R_{AT} = \{\{R_a\} | a \in AT\}$ は、各属性 $a \in AT$ の値の集合 V_a で定義された二項関係の集合 $\{R_a\}$ の集合。 $V = \bigcup_{a \in AT} V_a$ はすべての属性に関する値の集合であり、 $\rho: U \times AT \rightarrow V$ は対象 $x \in U$ の属性 $a \in AT$ での値 $\rho(x,a) \in V_a$ を表す関数である。属性の有限集合 AT を条件属性の有限集合 C と決定属性の有限集合 D に分割できるとき、 S を決定表と呼ぶ。

二項関係について、各属性の二項関係の集合 $\{R_a\}$ は等号や不等号など様々な二項関係を含むことが可能である。Pawlakのラフ集合では、すべての属性 $a \in AT$ について等号のみを含み、 $\{R_a\} = \{=\}$ となる。

3.2 相互関係決定表

既存のラフ集合論では、各対象が持つ属性値の比較が重要な役割を果たす一方で、その比較は各属性に対してそれぞれ個別に行われていた。そのため、2つの属性 a と b について、「 a と b の値が等しい」などの値の比較に基づく特徴は既存のラフ集合論では取り扱うことができなかった。

これに対して工藤ら[8]は、2つの属性間の相互関係を表現する関係性属性を新たな属性として加えることを提案した。前節での情報表 S の定義に基づき、相互関係決定表 S_{int} を以下の式(4)で定める[12]。

$$S_{int} = (U, AT_{int}, V \cup \{0,1\}, R_{int}, \rho_{int}) \quad (4)$$

ここで、対象の集合 U 及び属性値の集合 V は S と同一のものである。新たな二項関係の集合の集合 R_{int} は、以下の式(5)で定義される。

$$R_{int} = R_{AT} \cup \left\{ \{R_{a_i \times b_i}\} \mid \exists a_i, b_i \in C \right. \\ \left. \cup \{=\} \mid \text{For each } aRb \right\} \quad (5)$$

ここで、 $R_{a_i \times b_i}$ は属性 a_i, b_i の属性値の関係を表す集合 $R_{a_i \times b_i} = \{R_{a_i \times b_i}^1, \dots, R_{a_i \times b_i}^{n_i}\}$ であり、直積集合 $V_{a_i} \times V_{b_i}$ 上で定義された n_i 個の二項関係からなる集合である。

新たな属性の集合 AT_{int} は以下の式(6)で定義される。

$$AT_{int} = AT \cup \{aRb \mid \exists R \in R_{a \times b}, R(a,b) \neq \emptyset\} \quad (6)$$

ここで、 aRb を相互関係条件属性、または関係性属性と呼ぶ。

関係性マイニングと協調フィルタリングを用いた情報推薦手法

集合 $AT=CUD$ は元となる S の場合と同一である。 $R(a,b)$ は属性 a の属性値 $\rho(x,a)$ と属性 b の属性値 $\rho(x,b)$ との間に二項関係 R が成立する対象 $x \in U$ の集合であり、相互関係の支持集合と呼ぶ。 $R(a,b)$ は以下の式 (7) で定義される。

$$R(a,b) = \{x \in U | (\rho(x,a), \rho(x,b)) \in R\} \quad (7)$$

関数 $\rho_{int}: U \times AT_{int} \rightarrow V \cup \{0,1\}$ は式 (8) で定義され、対象 $x \in U$ の関係性属性を含むすべての属性 $c \in AT_{int}$ における属性値 $\rho_{int}(x,c) \in V_c$ を表す。

$$\rho_{int}(x,c) = \begin{cases} \rho(x,c), & \text{if } c \in AT, \\ 1, & c = aRb \text{ and } x \in R(a,b), \\ 0, & c = aRb \text{ and } x \notin R(a,b). \end{cases} \quad (8)$$

4. 提案手法

本節では、関係性マイニングの考え方をを用いて、アイテム間の比較に基づくユーザの選好パターンを考慮した推薦方法を提案する。対象ユーザの選好パターンを表す関係性属性を作成する際に、データセットに応じてアイテムに対する評価値の差の閾値を自動的に設定することで、より対象ユーザの特徴を反映した推薦を行う。

作成した関係性属性における値の一致度を用いてユーザ間の類似度を求め、2節の協調フィルタリングの考え方を適用し、以下の手順で推薦を行う。

4.1 関係性属性の作成

まず、対象ユーザ $x \in U$ の評価済みアイテムからアイテムの対 (a,b) を、対象ユーザによるアイテム b の評価値がアイテム a の評価値より閾値 Dif 以上大きくなるように n 個選択する。その際に、2個のアイテムに対する評価値の差の閾値をデータセットに応じて自動的に設定する。これは、対象ユーザの好みの違いが大きく表れている対を選ぶことを目的としている。値の差の閾値 Dif を、データセットの値域の最大値 $\max(Range)$ と最小値 $\min(Range)$ を用いて、以下の式 (9) で求める。

$$Dif = \frac{[\max(Range) - \min(Range)]}{2} \quad (9)$$

アイテムの対は重複を許さずに選択する。得られた対 (a,b) について、アイテム b の評価値がアイテム a の評価値より Dif 以上大きいことを表す二項関係 $a \leq_{Dif} b \subseteq V_a \times V_b$ を用いて関係性属性 $a \leq_{Dif} b$ を作成し、各ユーザ $u \in U$ に対する関係性属性 $a \leq_{Dif} b$ の値を式 (10) で定義する。

$$\rho_{int}(u, a \leq_{Dif} b) = \begin{cases} 0, & \text{if } \rho(u,b) - \rho(u,a) < Dif, \\ 1, & \rho(u,b) - \rho(u,a) \geq Dif, \\ *, & \rho(u,a) = * \vee \rho(u,b) = * \end{cases} \quad (10)$$

このとき、対 (a,b) の選び方から、対象ユーザの関係性属性の値はすべて1となる。

作成した各関係性属性は対象ユーザの選好パターンを表現する。ユーザ $u \in U$ の関係性属性 $a \leq_{Dif} b$ の値が1であるとき、

そのユーザ u は対象ユーザと同じ選好パターンを持ち、アイテム a よりアイテム b を (評価値の差が Dif 以上大きくなるほど) 強く好む。一方、 u の $a \leq_{Dif} b$ の値が0であるとき、 u は対象ユーザとは異なり、アイテム a とアイテム b に関して選好パターンを持たないか、あるいは逆の選好パターンを持つ。

4.2 予測値の計算および推薦アイテムの選定

対象ユーザ $x \in U$ と他のユーザ $u \in U - \{x\}$ の類似度として、対象ユーザと他のユーザの間で関係性属性の値が一致する割合を用いる。類似度は以下の式 (11) で定義する。

$$\begin{aligned} sim(x,u) &= \frac{|\{a \leq_{Dif} b \in AT_{int} | \rho(x, a \leq_{Dif} b) = \rho(u, a \leq_{Dif} b)\}|}{n} \quad (11) \end{aligned}$$

類似度の上位 k 人を最近傍ユーザとして用い、推薦を行う。アイテムに対する予測値は式 (2) で求める。予測値で上位 l 件のアイテムを推薦アイテムとして選定し、対象ユーザに提示する。

4.3 提案手法の使用例

例として、表1の評価値行列を用いて推薦を行うことを考える。対象ユーザをユーザ1とすると、ユーザ1が評価しているアイテムは $\{A,B,C,E\}$ である。評価値の値域の最大値は5、最小値は1なので、式 (9) より $Dif=2$ となり、二項関係は \leq_2 を用いる。対象ユーザの値が1になる関係性属性の例として $A \leq_2 B$ などが作成できる。表1から関係性属性を5個作成し、その属性値を求めた結果を表2に示す。

表2から、ユーザ1と他のユーザの類似度を式 (11) を用いて求めると、それぞれ $sim(1,2)=0.2$, $sim(1,3)=0.4$, $sim(1,4)=0.8$, $sim(1,5)=0$ となる。よって、類似度からユーザ1と最も類似しているユーザはユーザ4、次いでユーザ3、ユーザ2、ユーザ5、という順になる。ここでは最近傍ユーザとして2人選ぶとすると、 $N=\{3,4\}$ となる。

最近傍ユーザを用いて、ユーザ1のアイテムDとFの評価値を予測する。各ユーザの評価値平均は、それぞれ $\bar{r}_1=3.5$, $\bar{r}_3=3.4$, $\bar{r}_4=3.33$ であるため、ユーザ1のアイテムDに対する評価の予測値は、以下の式で求められる。

$$\begin{aligned} pred(1,D) &= 3.5 + \frac{(4-3.4) \times 0.4 + (5-3.33) \times 0.8}{0.4+0.8} \\ &= 4.81 \end{aligned}$$

同様に、アイテムFの予測値は、以下の式で求められる。

$$\begin{aligned} pred(1,F) &= 3.5 + \frac{(3-3.4) \times 0.4 + (3-3.33) \times 0.8}{0.4+0.8} \\ &= 3.14 \end{aligned}$$

表2 提案手法による選好パターン抽出例

ユーザ	$A \leq_2 B$	$A \leq_2 E$	$C \leq_2 A$	$C \leq_2 B$	$C \leq_2 E$
ユーザ1	1	1	1	1	1
ユーザ2	0	1	*	*	*
ユーザ3	*	0	1	*	1
ユーザ4	1	1	0	1	1
ユーザ5	*	*	*	0	*

予測値から、ユーザ1はアイテムFよりもアイテムDに対して高評価を与えると思われるため、アイテムDがユーザ1に対する推薦リストに加えられる。

4.4 協調フィルタリングとの比較

4.3節の使用例について、協調フィルタリングを用いた場合との比較を行う。表1について、協調フィルタリングによるユーザ1と他のユーザの類似度を式(1)により求めると、 $sim(1,2)=0.87$, $sim(1,3)=0.85$, $sim(1,4)=0.86$, $sim(1,5)=-0.24$ となる。このことから、ユーザ1と最も類似しているユーザはユーザ2、続いてユーザ4、ユーザ3、ユーザ5という順になる。提案手法と条件を同じにするため最近傍ユーザを2人選ぶと、 $N=\{2,4\}$ となる。

提案手法と同様に、ユーザ1のアイテムDとFに対する評価値の予測を行う。ユーザ2の評価値平均は $\bar{r}_2=3.6$ であるため、予測値は以下の式で求められる。

$$\begin{aligned} pred(1,D) &= 3.5 + \frac{(1-3.6) \times 0.87 + (5-3.33) \times 0.86}{0.87+0.86} \\ &= 3.01 \\ pred(1,F) &= 3.5 + \frac{(5-3.6) \times 0.87 + (3-3.33) \times 0.86}{0.87+0.86} \\ &= 4.04 \end{aligned}$$

予測値から、提案手法とは異なり、協調フィルタリングではユーザ1はアイテムDよりもアイテムFの方に高い評価を与えると推測される。よって、協調フィルタリングではアイテムFがユーザ1に対する推薦リストに加えられることになる。このように、提案手法と協調フィルタリングでは最近傍ユーザや推薦するアイテムが異なる可能性がある。

5. 実験

5.1 実験概要

提案手法の有用性を検証するために実験を行った。本実験ではMovieLens_100kデータセット[16]、MovieLens_1mデータセット[16]およびJesterデータセット[17]を用いた。各データセットでのユーザの人数、アイテムの個数および評価件数を表3に示す。

MovieLens_100kデータセットおよびMovieLens_1mデータセットは、映画に対する評価データであり、Web上で公開されているデータセットである(URL: <https://grouplens.org/datasets/movielens/>)。ユーザによるアイテム(映画)への評価は5段階評価で表され、1が最も低評価、5が最も高評価を意味する。どちらのデータセットも、すべてのユーザ

は少なくとも20件以上の映画に対して評価を行っている。各データセットで評価件数が最多のユーザは、MovieLens_100kデータセットでは737件、MovieLens_1mデータセットでは2314件の映画に対して評価を行っている。データの粗密の度合いを表す疎性(*Sparseness*)を式(12)で定義すると(値が1に近いほどデータセットが疎であることを表す)、MovieLens_100kデータセットでは $Sparseness=0.937$ 、MovieLens_1mデータセットでは $Sparseness=0.958$ となり、どちらも非常に疎性の高いデータセットである。このことから、2種類のMovieLensデータセットでは、ユーザの大半はほとんどの映画を未評価であることがわかる。

$$Sparseness = 1 - \frac{\text{評価件数}}{\text{ユーザ数} \times \text{アイテム数}} \quad (12)$$

Jesterデータセットは100件のジョークに対するユーザの評価データであり、Web上で公開されているデータセットである(URL: <http://eigentaste.berkeley.edu/dataset/>)。ユーザによるアイテム(ジョーク)への評価は-10から+10までの実数値で表され、値が大きいくほど高評価であることを意味する。Jesterデータセットは3分割された状態で公開されており、本研究では24983人のユーザによる評価データであるJester_dataset_1_1データセットを用いた。このデータセットでは、すべてのユーザは少なくとも36件以上のジョークに対して評価を行っており、その中で7200人のユーザは100件すべてのジョークに対して評価を行っている。データセットの疎性は $Sparseness=0.275$ である。よって、Jesterデータセットはアイテム数が非常に少なく疎性が低いデータセットであり、2種類のMovieLensデータセットとは性質が異なるデータセットである。

本実験ではLeave-one-out法を用いた。ユーザ集合から1人を対象ユーザとして選択し、4.1節の方法で対象ユーザの選好パターンを表す関係性属性を作成し、最近傍ユーザを求めた。データセット毎の値の閾値は、式(9)よりMovieLens_100kデータセットおよびMovieLens_1mデータセットでは $Dif=2$ を、Jesterデータセットでは $Dif=10$ を用いた。更に、対象ユーザの評価済みアイテムの内、関係性属性の作成に用いなかったアイテムすべてを未評価アイテムとみなし、それぞれの予測値を求めた。比較のため、関係性属性の作成に用いたアイテムのみを計算に用いて、ユーザベースの協調フィルタリングによる最近傍ユーザを求め、提案手法の場合と同様に、未評価とみなした各アイテムの予測値を求めた。このとき、関係性属性を規定の個数作成することができなかったユーザに関しては、計算からは除いている。

実験結果の評価指標として、*MAE*と*Coverage*を用いた。*MAE*は推薦精度の評価指標であり、式(13)で定義される。

$$MAE = \frac{1}{|U|} \sum_{u \in U} \left(\frac{1}{|O_u|} \sum_{i \in O_u} |p_{u,i} - r_{u,i}| \right) \quad (13)$$

ここで、 $p_{u,i}$ はアイテムの予測値、 $r_{u,i}$ は実際のアイテムの評価値、 O_u は対象ユーザの評価済みアイテムの内、関係性属

表3 用いたデータセットの詳細

データセット	ユーザ数	アイテム数	評価件数
MovieLens_100k	943	1682	100000
MovieLens_1m	6040	3952	1000209
Jester	24983	100	1810455

関係性マイニングと協調フィルタリングを用いた情報推薦手法

性の作成に用いられなかったアイテムの集合を表す。MAEの値が小さいほど、評価の予測値と実際の評価値との誤差が小さいことを表す。そのため、MAEの値が小さい推薦手法では、予測値の上位 l 件から推薦されたアイテムに対して、対象ユーザは高評価を与えることが期待されるため、精度のよい推薦を行えることが期待できる。

また、Coverageは推薦の多様性を評価する指標の一種であり、以下の式(14)で定義される。

$$Coverage = \frac{1}{|U|} \sum_{u \in U} \left(100 \times \frac{|I_u^c \cap Z_u|}{|I_u^c|} \right) \quad (14)$$

ここで、 I_u^c は対象ユーザの未評価アイテムの集合、 Z_u は最近傍ユーザが評価を与えているアイテムの集合を表す。Coverageは対象ユーザが未評価であるアイテムの中で、対象ユーザの最近傍ユーザが何等かの評価を与えているアイテムのパーセンテージの平均値であるため、この値が大きいほど、最近傍ユーザの情報を用いて予測値を求められる対象ユーザの未評価アイテムの割合が多いことを表す。よって、評価指標Coverageの値が大きい推薦手法では、対象ユーザが未評価である多くのアイテムの中から、対象ユーザが高評価を与えることが予測されるアイテムを多数見出すことが可能になるため、対象ユーザに対して多様な推薦が行えることが期待できる。

本実験では、これら2種類の評価指標について10回試行の平均を求めた。

更に、MovieLens_100kデータセットに対しては、適合率(Precision)と再現率(Recall)による評価も行った。適合率は以下の式(15)で、再現率は式(16)でそれぞれ定義される。

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

$$Recall = \frac{TP}{TP + TN} \quad (16)$$

ここで、TPは対象ユーザに対して推薦されたアイテムの中で、対象ユーザが実際に高評価を与えているアイテムの個数、FPは対象ユーザに対して推薦されたアイテムの中で、対象ユーザが低評価を与えているかまたは未評価であるアイテムの個数、TNは対象ユーザに対して推薦されなかったアイテムの中で、対象ユーザが高評価を与えているアイテムの個数である。適合率の値が高いほど、対象ユーザに対して推薦されたアイテムに対して、対象ユーザは高評価を与える確率が高くなるため、対象ユーザに対して精度のよい推薦を行えていることとなる。また、再現率の値が高いほど、対象ユーザが高評価を与えるアイテムを網羅的に推薦できていることとなる。一般的に、適合率と再現率はトレードオフの関係にある。

本実験では、MovieLens_100kデータセットにおける全ユーザについて適合率と再現率を求めてそれぞれの平均値を算出する試行を10回行い、10回試行の平均を求めた。

5.2 実験結果・考察

各データセットについて、提案手法と協調フィルタリング(CF)それぞれを用いて最近傍ユーザの数 k を20人、30人、40人、50人と変化させたときのMAEとCoverageを求めた。

また、MovieLens_100kデータセットおよびMovieLens_1mデータセットでは作成する関係性属性を10個と20個の2つの場合について求めた。Jesterデータセットではアイテムの個数を考慮し、関係性属性の個数を5個と10個とした。

MovieLens_100kデータセットで提案手法と協調フィルタリングを比較した結果を図1および図2に示す。図1および図2において、横軸は最近傍となるユーザの人数を、縦軸は各指標の値をそれぞれ表す。凡例にあるカッコつきの数字は、提案手法において作成した関係性属性の個数と、その比較対象となる協調フィルタリングを表す。同様に、MovieLens_1mデータセットで比較した結果を図3および図4に、Jesterデータセットで比較した結果を図5および図6にそれぞれ示す。

MovieLens_100kデータセットにおいて提案手法と協調フィルタリングを比較すると、図1から、提案手法はMAEの値が協調フィルタリングとほぼ同等である。これは対象ユーザの選好パターンとして、値の差を考慮した関係性属性を作成したことで、対象ユーザの好みの特徴を的確に捉えることができたためと考えられる。

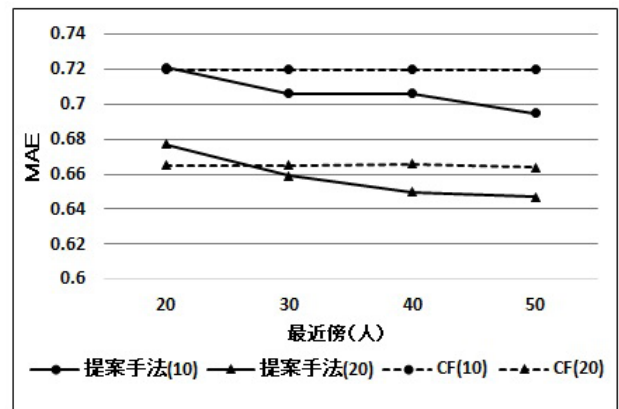


図1 MovieLens_100kでのMAEの比較

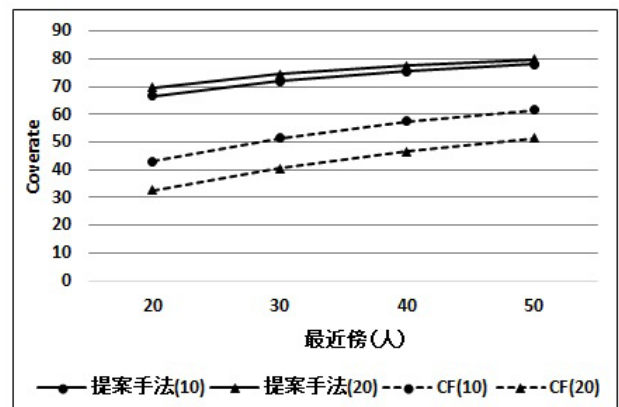


図2 MovieLens_100kでのCoverageの比較

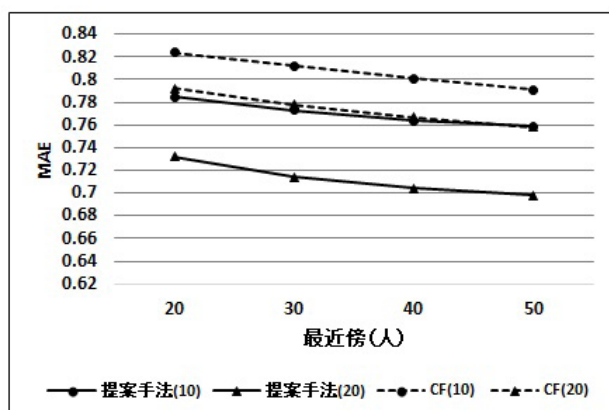


図3 MovieLens_1mでのMAEの比較

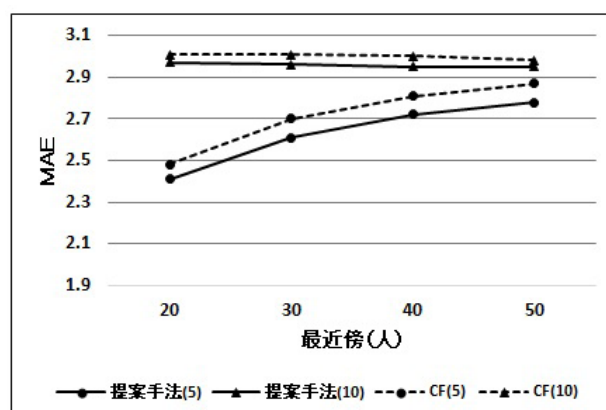


図5 JesterでのMAEの比較

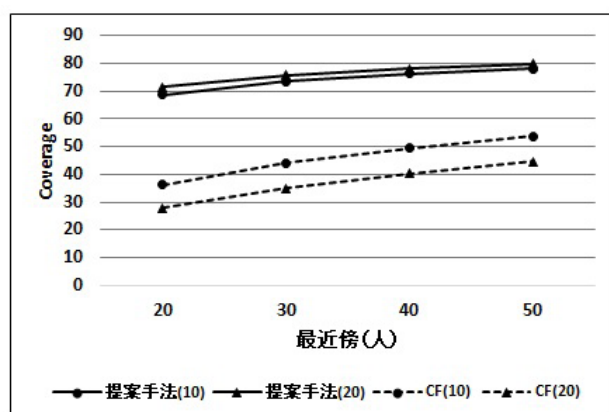


図4 MovieLens_1mでのCoverageの比較

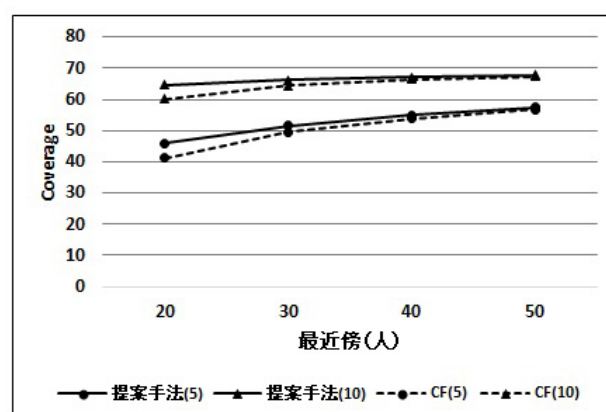


図6 JesterでのCoverageの比較

また、図2から、Coverageに関しては提案手法の方が協調フィルタリングと比較して高くなっている。このことから、提案手法ではより多様性に富んだ推薦が可能であると判断できる。これは、提案手法では関係性属性で表されているユーザの選好パターンについて、その一致度を求めることでユーザ間の類似度を求めているため、4.3節および4.4節におけるユーザ3のように、評価値を用いて類似度を直接求めた場合は最近傍ユーザとはならないが、選好パターンが似ていることで、提案手法では最近傍ユーザとみなされるユーザが多いためと考えられる。

図3および図4より、MovieLens_1mデータセットで提案手法は、MovieLens_100kデータセットの場合と同程度の推薦の多様性を保ちつつ、推薦精度では協調フィルタリングを上回った。また、最近傍ユーザの人数が増加するほど、推薦精度が改善される傾向が見られた。この理由として、値の差を考慮した関係性属性による選好パターンが、対象ユーザの好みの傾向を明確にすることにより、MovieLens_100kデータセットよりユーザの人数およびアイテム件数が共に増加したデータセットから、好みの傾向が似たユーザを的確に最近傍ユーザとして選定できていることが考えられる。

一方で、図5および図6から、Jesterデータセットでは提案手法と協調フィルタリングの間で指標に大きな差が見られない。また、関係性属性を5個とした場合に、提案手法と協調フィ

ルタリングの両方で、近傍ユーザ数が増加するほどMAEの値が悪化している。この原因として以下の2点が考えられる。

1. JesterデータセットはMovieLensデータセットとは異なり疎性が低いデータセットであること。
2. 提案手法では関係性属性5個、協調フィルタリングでは最大10個のアイテムの評価値のみを用いるため、類似度の測定に用いる情報が少ないこと。

以上の原因により、最近傍ユーザの人数が増加するほど、類似度評価では偶然的に対象ユーザと類似したものの、実験では未評価アイテムとして扱われたアイテム群については、対象ユーザと結果的にあまり類似しないユーザが最近傍に含まれたためと推測される。

最後に、MovieLens_100kデータセットに提案手法と協調フィルタリングそれぞれを用いて、推薦アイテム数 l を $l=30,50,70,90$ と変化させた時の適合率と再現率を求めた。この実験では、ユーザによる評価値が4以上のアイテムを高評価アイテム、3以下のアイテムを低評価アイテムとした。なお、最近傍ユーザの人数 k は、図1および図2の結果を踏まえ、MAEの値が最も小さく、かつCoverageの値が最も大きくなった $k=50$ で固定した。作成する関係性属性の個数について、10個と20個の2つの場合で実験を行った。

結果を図7および図8に示す。図7および図8において、横軸は推薦されたアイテムの個数を、縦軸は各指標の値をそ

関係性マイニングと協調フィルタリングを用いた情報推薦手法

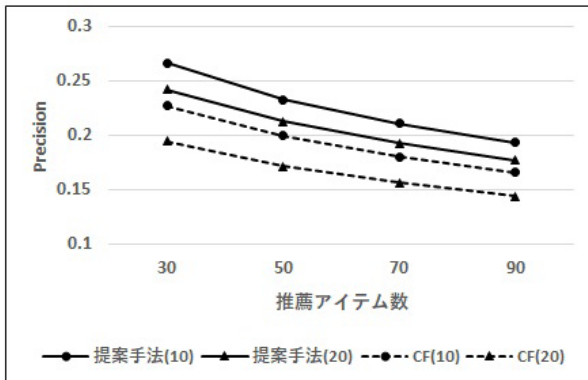


図7 MovieLens_100kでの適合率の比較

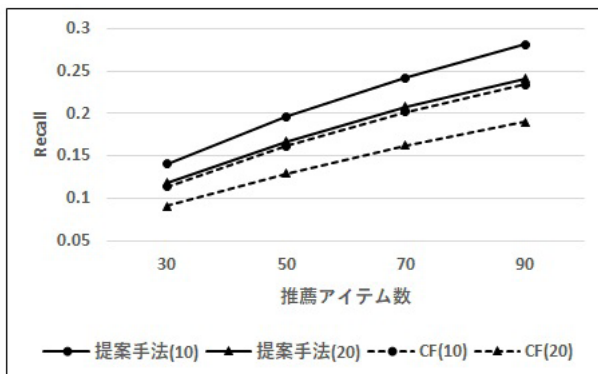


図8 MovieLens_100kでの再現率の比較

それぞれ表す。図1～図6と同様に、凡例にあるカッコつきの数字は、提案手法において作成した関係性属性の個数と、その比較対象となる協調フィルタリングを表す。

提案手法と協調フィルタリングを比較すると、すべての提案アイテム数で、提案手法による適合率および再現率は協調フィルタリングによる適合率および再現率を0.03～0.05程度上回った。これにより、提案手法による推薦内容は、少なくとも協調フィルタリングによる推薦内容と同程度以上の精度と網羅性を有することが示唆された。この要因として、MAEによる比較でも述べたように、対象ユーザの選好パターンとして値の差を考慮した関係性属性を作成したことで、対象ユーザの好みの特徴を的確に捉えることができたため、対象ユーザと好みの傾向が似た最近傍ユーザによる予測を用いて、適切なアイテムが推薦されたためと考えられる。

また、関係性属性の個数が10個の場合と20個の場合を比較すると、MAEやCoverageの場合とは異なり、関係性属性の個数が増加した20個の場合の方が、提案手法および比較対象の協調フィルタリングの両方で、適合率と再現率が共に悪化している。この原因として、対象ユーザの選好パターンとして作成する関係性属性の個数が増加することに伴い、関係性属性に使用するアイテムの個数、特に対象ユーザの高評価アイテムの個数が増加していることが考えられる。関係性属性に使用したアイテムは推薦の対象外としているため、関係性属性の個数が増加するほど、本実験において各対象ユーザに推薦される高評価アイテムの個数（式(15)および式(16)

におけるTP)は減少する。一方、実験で推薦されるアイテムの個数(TP+FP)は各実験で固定されており、各対象ユーザが高評価を与えているアイテムの個数(TP+TN)はユーザ毎に定まる。そのため、関係性属性の個数が10個から20個に増加したことで、適合率および再現率の定義式における分子の値のみ減少した結果、適合率および再現率の値が悪化する現象が起きたと考えられる。

6. まとめ

本論文では、ユーザベース協調フィルタリングの枠組みで推薦の精度と多様性を両立させる方策の1つとして、推薦を受ける対象ユーザによる評価値の差が大きいアイテムの対に着目し、これを対象ユーザの選好パターンとして明示的に選定し他ユーザと比較することを通じて、ユーザベース協調フィルタリングに対象ユーザの感性的な価値判断の傾向を明示的に反映させる手法を提案した。提案手法では、使用するデータセットに応じて評価値の差の閾値を自動的に設定することで、対象ユーザの選好パターンを関係性マイニングにおける関係性属性として明示的に抽出する。更に、選好パターンを通じて対象ユーザと他のユーザを比較することで、対象ユーザと好みの傾向が近いユーザを最近傍ユーザとして選定する。

提案手法を協調フィルタリングと比較した結果、提案手法は疎性の高いデータセットにおいて協調フィルタリングと同等以上の精度を持ちながら、多様性に富んだ推薦を行えることが確認された。一方で、疎性の低いJesterデータセットでは、協調フィルタリングとの性能の差は確認されなかった。

今後の課題として、提案手法で使用している、値の差の閾値、作成する関係性属性の個数、最近傍ユーザの人数などのパラメータと、データセットの疎性との関連性について詳細を調査する必要がある。また、推薦精度と推薦の多様性の両立を目指す他手法との比較、他のデータセットを用いた比較などが挙げられる。

更に、提案手法の要点である、対象ユーザの好みの違いを選好パターンとして抽出・比較する考え方は、情報推薦システムに限らず、ユーザの感性的評価を扱う様々な分野で使用できる可能性がある。そのため、提案手法の情報推薦以外の分野への応用可能性についても、今後の課題として検討する。

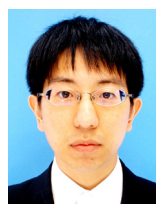
謝 辞

本研究は科研費基盤研究(C)(16K00365)の助成を受けたものである。

参 考 文 献

- [1] Herlocker, J. L., and Konstan, J. A.: An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. Information Retrieval, 5, pp.287-310, 2002.

- [2] Bradley, K., and Smyth, B.: Improving recommendation diversity, Proc. 12th Nat. Conf. Artif. Intell. Cogn. Sci. (AICS-01), Maynooth, Ireland, pp.75-84, 2001.
- [3] Vozalis, M. G., Markos, A. I., and Margaritis, K. G.: A Hybrid Approach for Improving Prediction Coverage of Collaborative Filtering, Proc. of AIAI 2009, pp.491-498, Springer, 2009.
- [4] Zhou, T., Kuscsik, Z., Liu, J.G., Medo, M., Wakeling, J. R., and Zhang, Y. C.: Solving the apparent diversity-accuracy dilemma of recommender systems, Proceedings of the National Academy of Sciences of the United States of America, 107(10), pp.4511-4515, 2010.
- [5] Gan, M. X., and Jiang, R.: Constructing a user similarity network to remove adverse influence of popular objects for personalized recommendation. Expert Systems with Application, 40, pp.4044-4053, 2013.
- [6] Gan, M. X., and Jiang, R.: Improving accuracy and diversity of personalized recommendation through power law adjustments of user similarities, Decision Support Systems, 55, pp.811-821, 2013.
- [7] Zhang, Z. P., Kudo, Y., and Murai, T.: Neighbor selection for user-based collaborative filtering using covering-based rough sets, Annals of Operations Research, 256(2), pp.359-374, 2017.
- [8] 工藤康生, 村井哲也: ラフ集合による関係性マイニングの構想, 第29回ファジィシステムシンポジウム講演論文集, pp.33-36, 2013.
- [9] 山脇淳一, 工藤康生, 村井哲也: 関係性マイニングと協調フィルタリングを用いた推薦手法の提案, 第12回日本感性工学会春季大会論文集, 1B-07, 2017.
- [10] 山脇淳一, 工藤康生, 村井哲也: 関係性マイニングと協調フィルタリングを用いた推薦手法の改良, 第19回日本感性工学会大会論文集, B26, 2017.
- [11] 田中克己, 角谷和俊 監訳: 情報推薦システム入門 一理論と実践一, 共立出版株式会社, pp.13-18, 2012.
- [12] Kudo, Y., and Murai, T.: A Review on Rough Set-Based Interrelationship Mining, Torra, V., Dahlbom, A., and Narukawa, Y. (eds.), Fuzzy Sets, Rough Sets, Multisets and Clustering, Springer, pp.257-273, 2017.
- [13] Pawlak, Z.: Rough Sets, International Journal of Computer and Information Science, 11, pp.341-356 1982.
- [14] Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publisher, 1991.
- [15] Yao, Y. Y., Zhou, B., and Chen, Y.: Interpreting Low and High Order Rules: A Granular Computing Approach, Proc. of RSEISP 2007, LNCS 4585, pp.371-380, 2007.
- [16] Harper, F.M. and Konstan, J.A.: The MovieLens Datasets: History and Context, ACM Transactions on Interactive Intelligent Systems (TiiS), 5(4), Article 19, 19 pages, DOI=<http://dx.doi.org/10.1145/2827872>, 2015.
- [17] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C.: Eigentaste: A Constant Time Collaborative Filtering Algorithm, Information Retrieval, 4(2), pp.133-151, 2001.



山脇 淳一 (正会員)

2016年 室蘭工業大学工学部情報電子工学系学科卒業。2018年 同大学大学院工学研究科情報電子工学系専攻修了。同年 株式会社OKI ソフトウェア入社。現在に至る。在学中は情報推薦システムに関する研究に従事。日本感性工学会では第12回春季大会優秀発表賞を受賞。



工藤 康生 (正会員)

1995年 北海道教育大学教育学部函館校総合科学課程卒業。1997年 北海道大学大学院工学研究科システム情報工学専攻博士前期課程修了。2000年 同博士後期課程修了, 博士(工学)。同年 室蘭工業大学SVBL博士研究員。2003年 同大学工学部情報工学科助手。以来, 助教, 准教授を経て, 2016年 同大学大学院工学研究科しくみ情報系領域教授, 現在に至る。IEEE, 日本感性工学会, 日本知能情報ファジィ学会, 人工知能学会各会員。



村井 哲也 (正会員)

1987年 北海道大学大学院情報工学専攻博士後期課程中途退学後, 札幌医科大学衛生短期大学部, 北海道教育大学函館校, 北海道大学大学院工学研究科・情報科学研究科を経て, 2016年より千歳科学技術大学理工学部情報システム工学科教授。1995年 博士(工学)(北海道大学)。日本感性工学会, 日本知能情報ファジィ学会, 人工知能学会各会員。