

## ユーザの嗜好の時系列性と先行性に着目した協調フィルタリング

川前 徳章<sup>†a)</sup>      坂野 鋭<sup>†</sup>      山田 武士<sup>†</sup>      上田 修功<sup>†</sup>

Collaborative Filtering Focusing on the Dynamics and Precedence of User Preference

Noriaki KAWAMAE<sup>†a)</sup>, Hitoshi SAKANNO<sup>†</sup>, Takeshi YAMADA<sup>†</sup>, and Naonori UEDA<sup>†</sup>

あらまし 本論文では、ユーザごとの嗜好の時系列性と、ユーザ間の行動の先行性に着目し、ユーザの最新の嗜好を反映できる協調フィルタリングを提案する。提案する協調フィルタリングでは、対象ユーザの購入アイテムを予測するために、過去の購入行動が対象ユーザと類似するだけでなく、先行している度合いの高いユーザの履歴を重視して選別し利用する。この先行性の尺度を、各時刻におけるユーザ自身の嗜好とユーザ間の購入時刻の差を反映して計算することで、各ユーザの最新の嗜好に合うアイテムの推薦が可能となった。オンラインストアの売上データを用いた実験によって、提案手法による推薦の予測精度は既存手法に比較し、比率で最大 51.5%の向上を示すことが分かり、その有効性を確認できた。

キーワード パーソナライズ, リコメンデーション, 購入予測, Innovator, 協調フィルタリング

### 1. ま え が き

近年、Web 上の多くのサービスにおいて、取り扱う情報やアイテム数の増加に伴い、それらの検索や予測が困難になっている。具体的には、ユーザが自身の嗜好に合うアイテムを検索すること、サービス提供者が多様化したユーザの嗜好に合うアイテムを予測すること、それぞれが負担となってきている。これらの負担を軽減するために、サービスのパーソナライズ、その実現技術である協調フィルタリングが注目されている。

しかしながら既存の協調フィルタリングには、ユーザの最新の嗜好に合うアイテムの推薦が難しいという問題がある。協調フィルタリングは「対象ユーザの嗜好に合うアイテムを見つけるための良い方法は嗜好が類似したユーザを見つけ、それらのユーザが好むアイテムを見つけることである」[1] という仮説に基づき、対象ユーザごとに「嗜好が類似した」ユーザを選別し、そのユーザの履歴を用いて推薦するアイテムを決定する。

既存手法では各ユーザの現時点までのアイテム購入履歴を一つのベクトルとしてとらえ、これらのベクトル間の類似度を計算し、履歴の類似度が高いユーザは嗜好が似ていると判断し履歴を選別する。既存の協調フィルタリングで用いる代表的な類似度には、cosine distance, correlation [2], や mean-squared difference [3] などがある。これらの類似度を単純に用いると、あるユーザが最近購入したアイテムとそうでないアイテムを同等に扱うために、時間変化を考慮して履歴を選別できない。その結果、既存手法により推薦されるアイテムは各ユーザの最新の嗜好に合わなくなる。

協調フィルタリングにおいて、最新の嗜好に合うアイテム推薦を実現するためには、ユーザや履歴の選別にユーザ行動の時間変化を反映させることが必要である。この点に着目して、いくつかの方法が提案されている。Ding [4] は時間の経過に応じて履歴に重みを付ける、時間減衰を用いることで、流行といったユーザ全体の嗜好の傾向を反映する協調フィルタリングを提案した。

一方、Pavlov [5] はユーザのアイテム購入の時間推移をモデル化し、推薦に利用する手法を提案した。このモデルを用いることで、嗜好の類似性に加えてアイテム購入順序を反映したアイテム推薦を実現できる。

Song [6] は、ユーザ間に伝搬する口コミや信頼性を

<sup>†</sup> 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所, 京都府

NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto-fu, 619-0237 Japan

a) E-mail: kawamae.noriaki@cslab.kecl.ntt.co.jp

モデル化する情報伝搬モデル [7] ~ [9] の考え方に基づき「あるユーザの購入行動に関する情報が、ユーザ間に伝搬し、他ユーザの購入行動に影響を与える」という仮説を主張し、この仮説に基づく Early Adoption Based Information Flow (以下, EABIF) を提案した。EABIF を用いることで、任意のアイテムを今後購入するユーザを予測できる。

協調フィルタリングが対象ユーザの最新の嗜好に合うアイテムを推薦できるためには、履歴の選別に「ユーザごとの嗜好」及び「ユーザ間の関係」の両方の時系列変化を反映させる必要がある。しかしながら、従来の研究にこれらを両方とも考慮した例は見られない。例えば Ding [4] は時間減衰を用いることで前者を考慮しているが、履歴を時刻により一様に選別するために、後者を考慮していない。また、アイテム間の時間推移を用いた Pavlov [5] も前者のみで後者を考慮していない。一方で、Song [6] は後者を考慮しているものの、前者を考慮していない。

本論文では履歴の選別にユーザごとの嗜好及びユーザ間の関係の時系列変化の両方を反映させた協調フィルタリングを提案する。そのために、我々はユーザの行動について「各ユーザは同様な嗜好をもつ先行ユーザと同様な行動をとり、先行性は時間変化する」という仮説を置く。ここで先行性とは、各ユーザと先行ユーザとの間を単純な購入時刻の前後関係 [6] ではなく、購入時刻の差で評価したものである。つまり提案手法ではユーザ間の関係を、行動の先行性とその時系列変化を考慮して定義し、これを先行度と呼び、履歴の選別に利用する。

先行度の計算に、ユーザ間の嗜好の先行性を反映するためにユーザ間の購入時刻の差を導入し、ユーザごとの嗜好の時系列性を反映するためにユーザごとの履歴の時間減衰を導入する。この先行度は筆者らが提案した Relative Innovator Degree (以下, RID) [10] からユーザの最新の嗜好を重視できるように発展させたものである。その結果、この先行度を用いた協調フィルタリングにより対象ユーザの最新の嗜好に合う履歴を用いた予測が可能になった。

以下、本論文では、2. においてユーザ間の嗜好の先行性と時系列性を説明し、それに基づいて先行度の定式化を行う。3. ではオンラインショップの売上データを用いて先行するユーザの履歴の有用性を確認し、先行度を用いた協調フィルタリング及び既存手法と予測精度の比較実験の結果を示す。4. では実験結果につい

て考察する。5. はまとめと今後の課題である。

## 2. 嗜好の先行性と時系列性に着目した協調フィルタリング

### 2.1 ユーザの嗜好と先行度

本節では、ユーザの嗜好の先行性と時系列性に着目することで、協調フィルタリングに有用な履歴を選別できるという我々の仮説を直感的に説明する。

購入履歴から観測できる各ユーザの購入アイテムとその購入時刻の 1 例を図 1 に示す。この図の横軸は時間の経過であり、各アイテムの位置は 4 名のユーザ  $u_a, u_1, u_2, u_3$  が該当アイテム  $i_{1-6}$  を購入した時刻を表している。ここでは、 $u_{1-3}$  の履歴を用いて Now 時点で  $u_a$  に対してアイテムを推薦する問題を考える。

既存の協調フィルタリングでは、ユーザ間の嗜好の類似性を該当ユーザ間の購入アイテムの類似性により計算する。この例の中で Now 以前では  $u_a$  と  $u_{1-3}$  が購入したアイテムは同じ  $i_{1-3}$  であるため、単純な類似度を用いた協調フィルタリングでは  $u_{1-3}$  の履歴を等価に扱う。したがって時刻 Now では  $u_{1-3}$  が購入した  $i_{4-6}$  のどれかを、あるいはすべてを  $u_a$  に推薦してもかまわないことになる。

しかし、購入時刻の違いに着目すると  $u_3$  がすべてのアイテムを  $u_a$  に遅れる形で購入しているのに対して  $u_1, u_2$  は三つのうち二つのアイテムを  $u_a$  より先に購入している。ここで我々の仮説を当てはめ、 $u_1, u_2$  は  $u_a$  の未来の嗜好、 $u_3$  は過去の嗜好を表すと考えると、 $u_a$  にとり  $u_1, u_2$  の履歴は  $u_3$  の履歴より有用であると考えられる。つまり、Now 時点で  $u_a$  に推薦するアイテムは  $i_6$  よりも、 $u_a$  より嗜好が先行する  $u_1,$

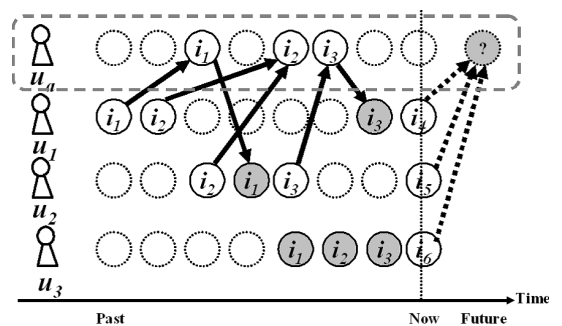


図 1 ユーザ間でのアイテム購入の先行関係  
Fig.1 Precedence relation between purchased items by the users.

$u_2$  が購入した  $i_4$  若しくは  $i_5$  が適切ということになる．すなわち、履歴の選別には嗜好の類似度に加えて先行性が有用であり、その先行性をユーザ間で共通するアイテムの購入時刻の差で与える．

ユーザ間で共通するアイテムの購入時刻の差が、該当ユーザの履歴の有用性と正の相関があることは次のように説明できる．嗜好が類似したユーザ集合の中には、アイテムのリリース前にその情報を知り、リリースと同時に購入するユーザが存在する．このような他に先駆けて購入するユーザは Innovator と呼ばれ、当該アイテムの属する分野に精通しており、新しい情報に敏感に反応すると考えられる [11]．したがって、Innovator の購入アイテムは嗜好が類似した一般ユーザにとって、他の一般ユーザの購入アイテムに比べ、より重要な情報となる．実際、CD や DVD 等の購入行動においても、常に他に先駆けてアイテムを購入するユーザの存在は確認できる．たとえ、映画「ブレア・ウィッチ・プロジェクト」のようにリリース後しばらくしてからヒットするケースであっても、Innovator は他者に先んじてこのようなアイテムを「発見」と考えられる．これらユーザの購入アイテムを、嗜好が類似した一般ユーザは遅れて購入している．図 1 の例では、 $u_1$  が  $u_a$  と共通するアイテムを購入した時刻は、 $u_2$  よりも早く、 $u_1$  は  $u_2$  よりもこの分野に精通しているといえる．つまり、 $u_1$  の購入アイテム  $i_4$  は、 $u_a$  にとって推薦価値が高い．そこで、ユーザ間の購入時刻の差を用いてユーザの先行性を与え、先行性を用いて推薦価値の高いアイテムを含むユーザの履歴を選別する．提案手法と [6] との違いはユーザ間の関係を時間的な順序ではなく時刻の差で与える点、及びユーザ間に情報伝達の存在を仮定しない点にある．ユーザ間の関係を時刻の差で与えることで、有用な履歴をもつ Innovator を正確に選別できるようになる．

一般に同一ユーザ組の中でも先行性はアイテムごとに異なり、相対的かつ動的な関係であることに注意する．この例では  $u_a$  にとり  $u_1, u_2$  は先行するユーザである．一方で、Now 時点近くでは  $u_a$  は  $u_1$  に先行して  $i_3$  を購入しており、 $u_a$  と  $u_1$  の先行性が入れ換った様子が観測できる．つまり将来的には  $u_a$  が  $u_1$  の先行ユーザとなる可能性がある．したがって、先行性の動的な変化を先行度に反映するために、購入時刻から現時点までのアイテムごとの時間経過も利用した．

これが提案する先行度の基本的なアイデアである．このアイデアの正当性をのちに実験によって検証する．

## 2.2 先行度の定義

本節では、前節での考察を反映する先行度の表式を与える．提案手法は、前節で述べたように、ユーザ間の先行性とその時系列変化の両方が履歴の選別に有用であるという仮説に基づく．つまり、先行度の計算にはユーザ間で共通するアイテムの数だけでなく共通アイテムごとの先行性とその時系列変化が必要である．そのため、先行度の計算に、ユーザ間で共通する購入アイテム数に加えて、嗜好の先行性及び時系列性を表現した以下の変数を導入する．

嗜好の先行性：ユーザ間の嗜好に基づく行動の先行性を該当ユーザ間で共通するアイテムの購入時刻の差を用いて表現する．アイテム  $i$  について  $u_b$  の  $u_a$  に対する先行の度合を  $r(i; b, a)$  で与える．

嗜好の時系列性：ユーザごとの嗜好の新鮮さをアイテム購入時刻を起点としたアイテム推薦時刻までの時間減衰を用いて表現する． $u_a$  の  $i$  の購入動機となった嗜好の時間減衰の係数を  $w(i; a)$  で与える．

先行性はユーザ間で共通するアイテムごとに、時系列性はユーザの履歴ごとに観測される．以上の議論を踏まえて、アイテム推薦の対象ユーザ  $u_a$  から見たデータセット内の任意のユーザ  $u_b$  の先行度  $PID(u_b, u_a)$  を以下の式で定義する<sup>(注1)</sup>．

$$PID(u_b, u_a) = \sum_{i \in C_{ab}} r(i; b, a) \times w(i; a) \quad (1)$$

ここで  $C_{ab}$  は  $u_a$  と  $u_b$  で共通する購入アイテムの集合である．

嗜好の先行性を表現する  $r(i; b, a)$  は単純にはアイテム購入順位の差により記述できる．実際に、筆者らが提案した RID はこの差を直接的に用いた先行度であり、以下のように  $\bar{r}(i; b, a)$  を与え、協調フィルタリングの性能向上を実現している．

$$\bar{r}(i; b, a) = \begin{cases} \frac{o_{a,i} - o_{b,i}}{N_i}, & \text{if } o_{b,i} \leq o_{a,i} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$RID(u_b, u_a) = \sum_{i \in C_{ab}} \bar{r}(i; b, a) \quad (3)$$

ここで  $o_{a,i}$  はアイテム  $i$  のリリース後から数えた  $u_a$  の購入順位、 $N_i$  はアイテム  $i$  を購入したユーザ数である．

(注1): 我々の先行度はマーケティング分野で用いられる Innovator とは異なり個人ごとに設定されるため personal innovator degree と称し、PID という記号を用いる．

本論文で提案する PID は実際の順位の代わりにあるアイテムの売れ行きの時間発展を表す確率密度関数を用いる．その結果，この尺度を確率として表現でき，検定などの統計的手法を適用できる．具体的にあるアイテムの売れ行きの時間発展を確率密度関数  $f(t)$  で記述できるとすると， $r(i; b, a)$  を以下のように与えることができる．

$$r(i; b, a) = \begin{cases} \int_{t_{b,i}}^{t_{a,i}} f(t) dt, & \text{if } t_{b,i} \leq t_{a,i} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

ここで  $t_{b,i}, t_{a,i}$  は  $u_b, u_a$  それぞれがアイテム  $i$  を購入した時刻である． $f(t)$  が不定積分可能な場合には累積分布関数  $F(t) = \int f(t) dt$  を用いて記述してもよい．

確率密度関数  $f(t)$  には，協調フィルタリングの適用サービスや対象にするアイテムごとに適した関数を用いることができる．

EC サイト等で流行性のあるアイテムを扱う場合には指数分布が適している．指数分布の確率密度関数は  $f(t; T) = \frac{1}{T} \exp(-\frac{t}{T})$ ，累積分布関数は  $F(t; T) = 1 - \exp(-\frac{t}{T})$  と与えられる．ここで  $t$  はリリースからユーザに購入されるまでの経過日数で  $t \geq 0$  を満たし， $T$  は  $t$  の期待値である．指数分布ではリリースから経過時間に比例して購入時刻の差に対応する確率の値は小さくなる．したがって，指数分布を用いた  $r(i; b, a)$  は，アイテムのリリースから購入までの経過時間を反映した確率的な意味をもつ．

本論文の実験で利用する，楽曲やビデオなどの購入ユーザ数の分布は，図 2 に示すように，ほぼ直線に従っているので，指数分布で近似してもさほど不自然ではない．なお，ここで用いたデータの詳細については，3.1 で詳しく説明する．図 2 の横軸はアイテムがリリースされてからの購入されるまでの経過日数，縦軸は経過日数ごとにアイテムを購入したユーザ数の分布の対数である．この図より購入したユーザ数の分布の対数を片対数グラフで直線に近似できること，すなわち，アイテムのリリース直後に購入ユーザ数のピークがあり，その数は指数的に減少することが確認できる．マーケティング関連の研究でも，経過日数に対する各アイテムを購入するユーザ数の分布の時系列変化が指数分布に従うことが知られている [11]．指数分布を用いて，式 (4) を以下のように記述できる．

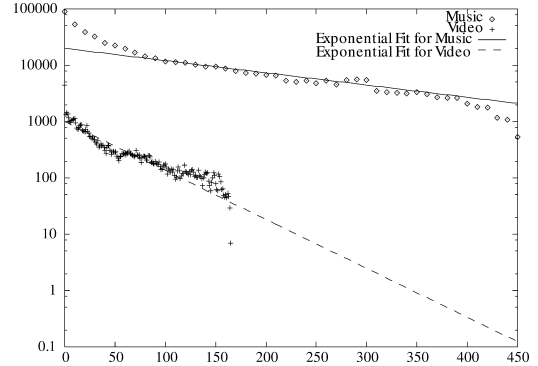


図 2 アイテムのリリース経過日数ごとの購入ユーザ数の分布

Fig. 2 Distribution of the number of users who purchased an item at times after its release.

$$r(i; b, a) = \begin{cases} \exp\left(-\frac{t_{b,i}-T_i}{\bar{t}_i}\right) - \exp\left(-\frac{t_{a,i}-T_i}{\bar{t}_i}\right), & \text{if } t_{b,i} \leq t_{a,i} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

ここで  $\bar{t}_i$  は  $i$  がリリースされてから購入されるまでの平均経過日数， $T_i$  は  $i$  のリリース日， $t_{a,i}$  及び  $t_{b,i}$  は式 (4) と同様である．

一般に，アイテム購入時点のユーザの嗜好は指数的に減衰すると考えられている [3], [11] ため，嗜好の時系列性を表現する  $w(i; a)$  は指数関数を用い以下のように記述する．

$$w(i; a) = \exp\left(-\frac{e_{a,i}}{\bar{e}_a}\right) \quad (6)$$

ここで  $e_{a,i}$  は  $u_a$  が  $i$  を購入してからアイテム推奨時点までの経過日数， $\bar{e}_a$  は  $u_a$  が購入した全アイテムについての  $e_{a,i}$  の平均である．Ding [4] の方法をそのまま適用すると，先行するユーザの履歴の重みが小さくなり，有用な履歴を選別できない問題がある．提案手法は [4] と異なり，指数関数をユーザ全体でなくユーザごとに与えたことで， $w(i; a)$  は  $i$  に関する  $u_a$  の平均的な購入パターンからのずれを評価する．ゆえに， $w(i; a)$  は  $u_a$  が購入したアイテム集合の中で，より最近購入したアイテムに対しより高い重みを与える．その結果，各ユーザの嗜好の新鮮さを重視した値をアイテムに与えることができる．

RID では嗜好の時系列性を考慮していなかったが，提案手法では  $w(i; a)$  を PID に導入することで，ユーザの最新の嗜好を重視できるように発展させた．その

結果、ユーザ間の嗜好の先行性及び時系列性を考慮して定義した PID により、各ユーザの最新の嗜好に適合した履歴の選別が可能となる。

### 2.3 先行度を用いたアイテムの推薦

提案手法では PID に比例する値でユーザの購入履歴を選別し、その履歴を用いて各ユーザへアイテムを推薦する。\$u\_a\$ が \$i\$ を購入する可能性 \$PP(i, u\_a)\$ を、式 (3) で定義した PID を用い、以下に与える式で求める。

$$PP(i, u_a) = \sum_b \delta(u_b, i) PID(u_b, u_a) \quad (7)$$

ここで、\$\delta(u\_b, i)\$ は \$u\_b\$ が \$i\$ を購入したか否かを表現する特性関数である。具体的には、\$u\_b\$ が \$i\$ を購入していれば \$\delta(u\_b, i) = 1\$、そうでなければ \$\delta(u\_b, i) = 0\$ となる。式 (7) より、\$i\$ を購入した PID の高いユーザが多いと、明らかにそのユーザ数に比例して \$PP(i, u\_a)\$ の値は高くなる。したがって、推薦対象となる全アイテムに対し \$PP(i, u\_a)\$ を計算すると、\$u\_a\$ に対して \$PP(i, u\_a)\$ の高い順に \$N\$ 個のアイテムを推薦する、す

なわち Top-\$N\$ のアイテム推薦を実現できる。

PID を用いた協調フィルタリングによるアイテム推薦のアルゴリズムを図 3 に示す。この一連の計算は購入履歴内のユーザ間で共通する購入アイテムとその購入時刻の比較により実行できる。ユーザ間の計算オーダはアイテム数を \$I\$ とすると \$O(I)\$ であり、従来の類似度を用いた場合と同程度である。またアイテム推薦時に、\$PP(i, u\_a)\$ の計算では PID が 0 以下の履歴は用いないため、計算量は小さくなる。したがって PID を用いた協調フィルタリングのアルゴリズムは、クリック履歴を用いた Web 検索等の他のパーソナライズサービスにも容易に実装可能である。

## 3. 実験

### 3.1 実験データ

本節では、本論文で利用するオンラインでの楽曲及びビデオのダウンロードサービスにおけるユーザの購入履歴、及び Netflix のレビュー履歴の詳細、及びそのデータを用いたシミュレーションの概要について述べる。

オンラインストアの購入履歴の詳細は以下のとおりである。楽曲ダウンロードの購入履歴は 2005/4/1 から 2006/7/31 までの 683,366 の購入セッション数であり、84,620 ユーザと 44,527 タイトルを含み、各購入履歴はユーザ ID、購入した曲のタイトルとその ID、アーティスト、購入日時及び価格から構成される。ビデオ視聴の購入履歴は 2005/9/1 から 2006/2/28 までの 66,087 の購入セッション数であり、7,537 ユーザと 4,064 タイトルを含み、各購入履歴はユーザ ID、購入したビデオのタイトルとその ID 及び購入日時から構成される。これらデータに含まれるアイテムは CD や Video (DVD) として店舗でのリリースと同時に該当サービスでも購入できる。購入データは評価データよりも疎であるため、各購入データにデフォルト投票 [1] を用いた。

協調フィルタリングのコンテストの評価データとして用いられる映画のレビュー履歴 Netflix [13] でも評価を行った。既存の協調フィルタリング [12] で用いられるデータセットはアイテムの購入履歴でなくユーザの評価データであり、1~5 までの五段階評価である。本論文で提案する協調フィルタリングでは、ユーザの評価でなく購入を予測する。そこで Netflix のデータを次のように利用した。まず、ユーザごとに評価が高い映画を購入したとみなす Herlocker [17] の方法を

---

```

Input: user logs
Output: item ranking $PP(i, u_a)$

calculate $\bar{t}_i$ and $\bar{e}_u$ on all items and users
$N_i$: Number of users who have purchased item $i$
$I$: Total number of unique items observed in user logs
$N_u$: Number of items purchased by user $u$
$U$: Total number of unique users observed in user logs
$T_0$: Present time (=the time to recommend items)

for $i = 1$ to $I$ do
    $\bar{t}_i \leftarrow \frac{1}{N_i} \sum_{j=1}^{N_i} (t_{j,i} - T_i) \dots(5)$
end for
for $u = 1$ to $U$ do
    $\bar{e}_u \leftarrow \frac{1}{N_u} \sum_{i=1}^{N_u} (T_0 - t_{u,i}) = \frac{1}{N_u} \sum_{i=1}^{N_u} e_{u,i} \dots(6)$
end for

calculate $PID(u_b, u_a)$ on each pair of $u_a$ and $u_b$
for $a = 1$ to $U$ do
    for $b = 1$ to $U$ do
        $PID(u_b, u_a) \leftarrow \sum_{i \in C_{ab}} w(i; a) \times r(i; b, a) \dots(3)$
    end for
end for

rank items for each target user $u_a$
for $a = 1$ to $U$ do
    $PP(i, u_a) \leftarrow \sum_b \delta(u_b, i) PID(u_b, u_a) \dots(7)$
    rank items in descending order of $PP(i, u_a)$
end for

```

---

図 3 PID を用いた推薦アルゴリズム

Fig. 3 Recommendation algorithm based on PID.

用いたものを Netflix (h) とした。この方法はユーザが 4 あるいは 5 と評価した映画をユーザが購入した映画とみなす。実際には、Netflix では評価の 96% が 4 あるいは 5 と大きく偏っている。そのため、この方法を用いた Netflix (h) は Netflix の元データとの違いがほとんどない。そこでユーザごとに評価の平均値を計算し、各ユーザの評価がそのユーザごとの平均より高い映画をそのユーザが購入した映画とみなしてデータを利用するものを Netflix (p) とし利用した。本論文で利用する Netflix のレビューデータの詳細は以下のとおりである。もとのデータは 1999/12/31 から 2005/12/31 までの 100,480,507 レビュー数からなり、480,189 ユーザと 17,770 タイトルを含む。各レビュー履歴はユーザ ID, タイトル ID, 各ユーザの評価とその評価日付より構成される。本論文では 2005/12/1 から 2005/12/31 までをテスト期間とし、2005/5/1 から 2005/11/31 までを学習期間とした。

次にこれらデータセットを検証と比較実験のために、学習期間とテスト期間に分けた結果を表 1 に示す。このようにデータを分けることで、学習期間の購入履歴を用いてテスト期間の購入アイテムを予測するシミュレーションが可能となる。このシミュレーションでは、協調フィルタリングにより楽曲をユーザに推薦する日は 2006 年の 7 月 1 日、ビデオでは 2006 年の 2 月 15 日、Netflix では 2005 年の 12 月 1 日とした。

3.2 嗜好の先行性の効果

本節では実験に先立ち、ユーザ間の嗜好の先行性に注目することで協調フィルタリングに有用な履歴を選別できるという提案手法の仮説を検証する。ここでは履歴の有用性を、適合率でなく再現率で評価した。すなわち、協調フィルタリングのシミュレーションの評価に、対象ユーザの購入アイテムを予測する「適合率」

を用いたのに対し、この検証には対象ユーザの購入アイテムを事前に購入したユーザを全ユーザから検索する「再現率」を用いた。

表 1 に示したデータを用い、履歴の有用性を検証する方法を説明する。まず、対象ユーザ  $u_a$  が学習期間に購入したアイテムを学習アイテム集合  $C_a$ 、テスト期間に購入したアイテムをテストアイテム集合  $D_a$  とする。この検証で  $C_a$  を  $u_a$  の購入履歴、 $D_a$  を  $u_a$  の将来の購入行動と仮定する。したがって、 $u_a$  の購入行動を予測するためには、 $D_a$  に含まれるアイテムを学習期間中に購入したユーザが必要である。すなわち  $C_b \cap D_a \neq \phi$  の条件を満たすユーザ  $u_b$  の購入履歴が  $u_a$  の購入行動の予測にとって必要で、結果的に有用となった履歴といえる。

ここで提案手法が用いる先行度の有用性を評価するために、以下の量を用いた。

$$\{C_b \triangleleft C_a\} : C_b \cap C_a \text{ のうち } u_b \text{ が } u_a \text{ より先に購入したアイテム集合}$$

次にこの量を用いて、 $C_a$  に対応する  $C_b$  の数を以下のように求めた。ここで  $C$  はデータ内で観測される全ユーザの履歴集合である。

$$\bullet C_a(k, m) = \{C_b \in C \mid |C_b \cap C_a| = k, |C_b| - |C_a| = m\} : \text{各 } C_a \text{ に対し共通アイテム数が } k \text{ 及び購入アイテム数の差が } m \text{ である履歴集合}$$

$$\bullet C_a^*(k, m) = \{C_b \in C \mid |C_b \cap C_a| = k, |C_b| - |C_a| = m, C_b \cap D_a \neq \phi\} : C_a(k, m) \text{ のうち } C_b \cap D_a \neq \phi \text{ である履歴集合}$$

$$\bullet C_a(k, l, m) = \{C_b \in C \mid |C_b \cap C_a| = k, |C_b \triangleleft C_a| = l, |C_b| - |C_a| = m\} : C_a(k, m) \text{ のうち } u_b \text{ が先に購入したアイテム数が } l \text{ である履歴集合}$$

$$\bullet C_a^*(k, l, m) = \{C_b \in C \mid |C_b \cap C_a| = k, |C_b \triangleleft C_a| = l, |C_b| - |C_a| = m, C_b \cap D_a \neq \phi\} : C_a(k, l, m) \text{ のうち}$$

表 1 実験に用いたデータセット  
Table 1 The data sets used in the experiments.

データの種類	楽曲	ビデオ	Netflix (o)	Netflix (h)	Netflix (p)
テスト期間	2006/7/1 以降に 1 曲でも購入したユーザの 2006/7/1 以降の購入履歴 (8,756 ユーザ/6,814 タイトル)	2006/2/15 以降に 1 タイトルでも購入したユーザの 2006/2/15 以降の購入履歴 (2,579 ユーザ/1,887 タイトル)	2005/12/1 以降に 1 タイトルでもレビューしたユーザの 2005/12/1 以降のレビュー履歴 (13,589 ユーザ/9,264 タイトル)	2005/12/1 以降に 1 タイトルでもレビューしたユーザの 2005/12/1 以降のレビュー履歴 (13,589 ユーザ/9,221 タイトル)	2005/12/1 以降に 1 タイトルでも購入したとみなせるユーザの 2005/12/1 以降のレビュー履歴 (13,589 ユーザ/5,365 タイトル)
学習期間	テスト期間に含まれるユーザの 2006/7/1 以前の購入履歴 (8,756 ユーザ/24,566 タイトル)	テスト期間に含まれるユーザの 2006/2/15 以前の購入履歴 (2,579 ユーザ/3,766 タイトル)	テスト期間に含まれるユーザの 2005/12/1 以前のレビュー履歴 (13,589 ユーザ/11,327 タイトル)	テスト期間に含まれるユーザの 2005/12/1 以前のレビュー履歴 (13,589 ユーザ/11,108 タイトル)	テスト期間に含まれるユーザの 2005/12/1 以前のレビュー履歴 (13,589 ユーザ/7,382 タイトル)

$C_b \cap D_a \neq \phi$  である履歴集合

最後にデータ内の全ユーザをそれぞれ対象ユーザとみなし、この合計を求める。既存の先行性を考慮しない選別手法で有用な履歴の含まれる割合  $r(k, m)$ 、及び先行性を考慮した選別手法で有用な履歴の含まれる割合  $r(k, l, m)$  をそれぞれ次のように求める。  $l$  のみの効果を確認するために、 $k$  及び  $m$  が一定という条件のもとで  $l$  の効果を比較した。

$$r(k, m) = \frac{\sum_a |C_a^*(k, m)|}{\sum_a |C_a(k, m)|},$$

$$r(k, l, m) = \frac{\sum_a |C_a^*(k, l, m)|}{\sum_a |C_a(k, l, m)|} \quad (8)$$

有用な履歴の含まれる割合が高くなれば、 $r(k, m)$  及び  $r(k, l, m)$  の値は 1 に近づく。また各  $k, m$  について  $r(k, l, m) > r(k, m)$  が成立すれば、先行性のあるユーザの履歴が協調フィルタリングに有用であることを確認できる。

図 4 に例として、共通アイテム数  $k = 2$  での購入アイテム数の差  $m$  に対する  $r(k, m)$  及び  $r(k, l, m)$  の変化の様子を示す。この図よりほぼすべての  $m$  について  $l$  の増加と  $r(k, l, m)$  の増加に正の相関が見られることから、有用な履歴の選別の条件に  $l$  が有効であることが分かる。多くの  $m$  について  $r(2, 2, m) > r(2, 1, m) > r(2, 0, m)$  の大小関係と  $r(2, 1, m) \simeq r(2, m)$  の関係が確認できる。つまり先行しているユーザの方が有用な履歴を含む割合が高いことが分かる。したがって、協調フィルタリングに先行性を利用することが有用であることが示唆される。

### 3.3 既存手法との比較実験

本節では、アイテム予測における提案手法と既存手法の比較実験の結果を示す。

本実験では、予測精度比較のベンチマークとして協

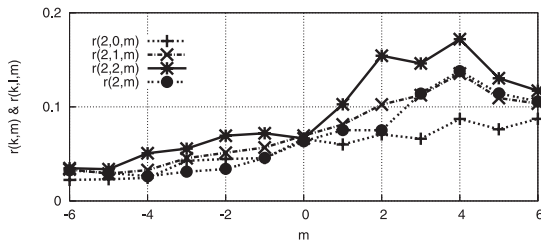


図 4  $k = 2$  における  $r(k, m)$  及び  $r(k, l, m)$  の変化 ( $x$  軸は  $m$ ,  $y$  軸は  $r(k, m)$  及び  $r(k, l, m)$  の値)

Fig. 4 The values of  $r(k, m)$  and  $r(k, l, m)$  with  $k = 2$  for different values of  $m$  and  $l$ .

調フィルタリングの代表的な手法である Pearson, Cosine [2], Item Based [14], bPLSA ( [15] の gPLSA に購入の有無を予測するためにガウス分布をベルヌーイ分布としたもの ), EABIF ( [6] の EABIF の時間推移を逆に使い、各アイテムでなく各ユーザに定義したもの ), 及び RID [10] を用いた。テスト期間の開始時点で人気の高い順にアイテムを推薦すると、全ユーザに共通のアイテムが推薦される。この方法をここでは Popular と呼び、比較対象とした。

ここで比較実験の概要を示す。アイテム推薦のシミュレーションをするために、表 1 に示したデータセットを更にランダムに  $K$  個のデータセットに分割した。

$K$  個に分割したうち 1 個のデータセットのテスト期間に含まれる全ユーザをそれぞれ対象ユーザとし、提案手法を用いて残りの  $K - 1$  個のデータセットからアイテムをランキングした。そのランキングに基づいて対象ユーザに対して上位  $N$  件のアイテムを提示した。これら提示したアイテムが対象ユーザのテスト期間中の購入履歴に含まれるかどうかを確認した。このシミュレーションは協調フィルタリングのシナリオに基づいた典型的なランキング方法である。

この方法に対する評価指標に、協調フィルタリングの予測精度を評価する共通の指標である Top  $N$  precision を使い、 $K = 10$  での各手法の予測精度を比較した。Top- $N$  リコメンデーションでは  $N$  を 1, 5, 10 と変化させた。この結果を表 2 に示す。Top-10 における推薦アイテムの被覆率と Gini 係数を表 3 に示す。推薦ユーザ/アイテムの被覆率は、テスト期間にアイテムを購入したユーザ数/購入されたアイテムのタイトル数のうち各手法が推薦可能なユーザ/タイトル数の比率である。Gini 係数は、手法ごとに推薦可能なアイテム集合の、テスト期間における購入ユーザ数の分布状況である。この係数は 0 から 1 までの値をとり、値が 0 に近いほどアイテムごとの購入ユーザ数の格差が少なく、1 に近いほど格差が大きいことを意味する。換言すれば、値が 1 に近い場合には特定のアイテムばかりが推薦される傾向が強くなり、ユーザごとの推薦アイテムの違いは小さくなる。すなわち推薦結果がパーソナライズされていないといえる。

表 2 の結果より全手法の中で PID, 次いで RID の精度が高く、協調フィルタリングの履歴の選別に先行性が有用であることが改めて示された。更に、PID が RID より予測精度が高いことから嗜好の時系列性の有用性も示された。実験に用いたデータセット内のアイ

表 2 楽曲, ビデオ及び Netflix データにおけるリコメンデーションの Top N による比較

Table 2 Top N precision of personalized recommendation with the different models on Music, Video and Netflix datasets.

Data set	top N	Popular	Cosine	Pearson	Item	bPLSA	EABIF	RID	PID
楽曲	1	6.50	6.90	6.91	6.10	5.98	13.12	17.79	<b>20.62</b>
	5	21.69	23.01	23.03	20.34	19.93	26.24	29.85	<b>33.97</b>
	10	33.12	32.25	31.47	29.84	27.85	34.18	37.25	<b>38.11</b>
ビデオ	1	7.84	8.31	8.31	7.35	7.20	15.62	19.72	<b>23.98</b>
	5	26.14	27.73	27.75	24.51	24.02	32.43	38.32	<b>39.97</b>
	10	32.27	34.23	34.26	30.26	29.65	37.19	41.33	<b>44.91</b>
Netflix (h)	1	5.88	5.12	5.32	4.76	5.23	6.81	7.53	<b>8.21</b>
	5	7.35	6.33	7.21	6.35	8.12	8.85	10.71	<b>11.59</b>
	10	9.82	7.56	8.46	7.22	10.21	11.32	14.43	<b>15.58</b>
Netflix (P)	1	6.37	5.83	6.01	5.26	6.68	7.72	8.53	<b>9.32</b>
	5	9.21	7.27	7.54	6.98	10.37	11.51	13.26	<b>14.56</b>
	10	10.32	8.32	9.57	8.51	12.86	13.44	16.79	<b>17.62</b>

表 3 楽曲, ビデオ及び Netflix データにおけるリコメンデーション結果の被覆率及び Gini 係数による比較

Table 3 Top N precision of personalized recommendation with the different models on Music, Video and Netflix datasets.

Data set	評価尺度	Popular	Cosine	Pearson	Item	bPLSA	EABIF	RID	PID
楽曲	推薦ユーザの被覆率	98.9	98.9	98.9	98.9	98.9	97.7	97.7	97.7
	推薦アイテムの被覆率	6.50	6.90	6.91	6.10	5.98	13.12	17.79	<b>20.62</b>
	Gini 係数	0.51	0.47	0.51	0.49	0.51	0.45	0.42	<b>0.39</b>
ビデオ	推薦ユーザの被覆率	99.2	99.2	99.2	99.2	99.2	98.6	98.6	98.6
	推薦アイテムの被覆率	8.43	8.87	8.72	12.31	11.56	18.55	26.39	<b>32.41</b>
	Gini 係数	0.46	0.42	0.47	0.43	0.46	0.21	0.29	<b>0.27</b>
Netflix (h)	推薦ユーザの被覆率	92.3	92.3	92.3	92.3	92.3	91.5	91.5	91.5
	推薦アイテムの被覆率	5.36	6.12	6.17	6.88	9.54	10.32	11.57	<b>12.11</b>
	Gini 係数	0.72	0.76	0.74	0.71	0.69	0.58	0.52	<b>0.51</b>
Netflix (P)	推薦ユーザの被覆率	90.9	90.9	90.9	90.9	90.9	90.3	90.3	90.3
	推薦アイテムの被覆率	5.88	5.93	5.85	7.21	6.35	11.1	14.21	<b>15.54</b>
	Gini 係数	0.67	0.62	0.63	0.58	0.61	0.52	0.48	<b>0.44</b>

テムは日用品でなく嗜好品である。これはユーザのアイテム購入の動機が需要よりも嗜好に依存することを意味する。そのために、対象ユーザが最近購入したアイテムは以前購入したアイテムより対象ユーザの最近の嗜好に合っていると考えられる。RID では嗜好の時系列性を区別していなかった。PID はアイテム購入時刻に伴う時間減衰を利用することで先行するユーザ群を対象ユーザの最近の嗜好との近さから選別し、対象ユーザへのランキングに反映した結果、予測精度を向上した。

また時系列変化の有用性はデータセットごとの予測精度の違いからも確認できる。実験データの節で述べたように、楽曲とビデオは購入データであり、Netflix は評価データである。一般に評価は購入の後に行われ、しかも購入と評価のタイムラグは固定的ではないため、評価時刻は購入時刻ほどユーザ間の先行性及び嗜好の時間変化を見るのに有用な情報でないと考えられる。そのようなデータ、Netflix (h) や Netflix (p) で

も先行性の効果は確認できるが、最も予測精度の高い PID でも、TOP-10 での精度は 20% を超えていない。一方、購入データである楽曲とビデオは、PID による予測精度は 40% 前後であり、他の手法に比較して精度の上昇が高いことも確認できる。したがって、嗜好の時間変化があるデータに対し、その変化を考慮している提案手法はより効果的であることが明らかになった。

表 3 の結果より全手法の中で PID を用いた場合の被覆率は最も高く、Gini 係数は最も 0 に近い値を示した。このことは、PID により多くのアイテムを偏りなく推薦できることを示している。一方で、ユーザ間の時系列的な関係を考慮していない従来手法は、被覆率は小さく、Gini 係数は 1 に近い値を示した。その理由は、これらの手法は先行度が小さい他ユーザの履歴まで用いるために、一部の人気アイテムを推薦結果の上位にランキングする傾向が強くなったためと考えられる。ユーザ間の関係の時間変化を反映した手法は、ユーザの被覆率は若干減少している。しかし、これら

減少したユーザは他ユーザの履歴を必要としない先行ユーザである．先行性を反映していない他手法ではこのユーザに推薦できるものの，その予測精度はほとんど0だった．

#### 4. 考 察

表2の結果より，提案手法の有効性を推薦結果に含まれるアイテムの特徴から考察する．

Popularの予測精度は時系列を考慮しない典型的な協調フィルタリング手法とほぼ同程度である．この結果は実験データに対し最近の人気アイテムに着目し推薦するだけでも効果があり，アイテムの時系列性を考慮することが有用であることを示唆している．

Itemとはアイテム間の類似性に基づく手法であり，この中では最も低い精度を示した．その理由は，購入履歴に含まれる期間すなわち学習期間に，最新のアイテムは古いアイテムほど購入されていないためである．つまり，購入履歴に含まれる最新アイテムベクトルは古いアイテムと比較して多くの要素が0となりスパースとなる．したがって，最新アイテムより古いアイテムの方が類似性が高くなり，上位にランキングされるようになる．その結果，Itemに基づく推薦結果がユーザの最近の嗜好と外れるために，予測精度が低下した．

ユーザ間の関係の時間変化を反映する手法とそうでない手法を比較すると，EABIF，RID及びPIDの精度はCosine及びPearsonの精度より高い．この結果より協調フィルタリングによる推薦に，我々が仮定したようにユーザ間の先行性に着目して履歴を選別することが有用であることを確認できる．

先に述べたように，ここでのEABIFは「アイテムごとのユーザ予測」であったものを「ユーザごとのアイテム予測」ができるように，定義を変えて利用している．つまり，EABIFではユーザ間の関係を時間的な前後関係で与えるのに対し，RID及びPIDでは時刻の差に基づいて詳細に与えている点が異なる．更にPIDでは先行性の計算に購入時刻の時間減衰を導入している．この時間減衰をアイテムのリリース日を起点とし，アイテムを重み付けできる点で，提案手法は他の手法に比較して有利である．その結果，提案手法は嗜好の先行性と時系列性をユーザ間の関係に反映し，協調フィルタリングの予測精度の向上を実現した．

協調フィルタリングの評価には推薦結果の精度だけでなく，推薦アイテムのセレンディピティも重要である[17]．推薦アイテムのセレンディピティとは，(1)「目

新しさ，予見のできなさそして意外性」[16]，(2)「グループ向けの推薦には現れないが，個人向けの推薦には現れるもの」という概念である[17]．提案手法は先行性に着目し，対象ユーザと共通するアイテムを先に購入したユーザのみの履歴を選別し利用することで，対象ユーザに適したアイテムを確実に推薦できるようになり，アイテム推薦の予測精度が向上した．同時に，これら選別したユーザの中でも，対象ユーザと共通するアイテムをリリース直後に購入したユーザの履歴をより重視して利用することで，これら先行ユーザの購入アイテムにより高い重みを与えた．その結果，先行ユーザの購入アイテムが上位にランキングされるようになり，アイテム被覆率が上昇した．アイテム被覆率の上昇は，各ユーザごとに異なるアイテムが推薦されていることを意味し，Gini係数は減少した．提案手法が精度だけでなくセレンディピティも達成しているかについての検討は今後の課題である．

#### 5. む す び

本論文ではユーザ間の先行性の尺度であるPIDとそれに基づく協調フィルタリングを提案し，実験によりその有効性を確認した．

嗜好の先行性と時系列性に着目しユーザの履歴を選別することがアイテム推薦の予測精度の向上に有用であることを実験により確認できた．

今後は，この先行度を統計的な見地から検証し，より普遍的なモデルへ拡張する．更に，購入がリピートされるアイテムにも適用できるモデルへ拡張する．

#### 文 献

- [1] J.S. Breese, D. Heckerman, and C.M. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," 14th Annual Conference on Uncertainty in Artificial Intelligence, pp.43-52, July 1998.
- [2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "An open architecture for collaborative filtering of netnews," CSCW, pp.175-186, ACM Press, 1994.
- [3] D. Widyantoro, T. Ioerger, and J. Yen, "Learning user interest dynamics with a three-descriptor representation," Journal of the American Society for Information Science and Technology, vol.52, no.3, pp.212-225, 2001.
- [4] Y. Ding and X. Li, "Time weight collaborative filtering," Proc. 14th ACM International Conference on Information and Knowledge Management (CIKM), pp.485-492, May 2005.
- [5] D. Pavlov and D. Pennock, "A maximum entropy ap-

proach to collaborative filtering in dynamic, sparse, high-dimensional domains,” *Advances in Neural Information Processing*, vol.15, pp.1441–1448, 2002.

- [6] X. Song, C.Y. Lin, B. Tseng, and M.T. Sun, “Personalized recommendation driven by information flow,” *ACM SIGIR*, pp.509–516, 2006.
- [7] M. Richardson and P. Domingos, “Mining knowledge-sharing sites for viral marketing,” *ACM SIGKDD*, pp.61–70, 2002.
- [8] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” *ACM EC*, pp.158–167, 2000.
- [9] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, “Propagation of trust and distrust,” *WWW*, pp.403–412, 2004.
- [10] 川前徳章, 山田武士, 上田修功, “Relative Innovator の発見によるパーソナライズ手法の提案,” *情報科学技術レターズ*, vol.6, pp.99–102, 2007.
- [11] E.M. Rogers, *Diffusion of Innovations*, The Free Press, New York, 1995.
- [12] M.R. McLaughlin and J.L. Herlocker, “A collaborative filtering algorithm and evaluation metric that accurately model the user experience,” *ACM SIGIR*, pp.329–336, 2004.
- [13] <http://www.netflixprize.com/>
- [14] B. Sarwa, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” *WWW*, pp.285–295, 2001.
- [15] T. Hofmann, “Collaborative filtering via gaussian probabilistic latent semantic analysis,” *ACM SIGIR*, pp.259–266, 2003.
- [16] 神鷹敏弘, “推薦システムのアルゴリズム (3),” *人工知能誌*, vol.23, no.2, pp.248–263, 2008.
- [17] J.L. Herlocker, J. Konstan, L. Terveen, and J. Riedl, “Evaluating collaborative filtering recommender systems,” *ACM Trans. Inf. Syst.*, vol.22, no.1, pp.5–53, 2004.
- [18] U. Shardanand and P. Maes, “Social information filtering: Algorithms for automating “word of mouth,”” *CHI*, pp.210–217, May 1995.

(平成 20 年 10 月 8 日受付, 21 年 1 月 17 日再受付)



川前 徳章 (正員)

1996 早大・理工卒。1998 東京大学大学院工学系研究科修士課程了。2002 同大学院工学系研究科博士後期課程了。博士(工学)。同年, NTT 入社。現在 NTT コミュニケーション科学基礎研究所研究員。情報検索, 統計的機械学習及びマーケティング

の研究に従事。FIT2007 船井ベストペーパー賞受賞。



坂野 鋭 (正員)

昭 63 中大・理工・物理卒。平 2 埼玉大学大学院理学研究科修士課程了。同年, NTT データ通信(当時)入社。平 20 より NTT コミュニケーション科学基礎研究所主任研究員。博士(工学)。統計的パターン認識と学習理論の研究に従事。日本物理学会, 情報処理学会, 画像電子学会各会員。



山田 武士 (正員)

1988 東大・理・数学卒。同年 NTT 入社。1996 より 1 年間英国コベントリー大学客員研究員。現在, NTT コミュニケーション科学基礎研究所創発環境研究グループリーダー。主として機械学習, メタヒューリスティクスによる組合せ最適化等の研究に従事。博士(情報学)。情報処理学会, ACM, IEEE 各会員。



上田 修功 (正員)

1982 阪大・工・通信工学卒。1984 同大学院修士課程了。工博。同年 NTT 入社。1993 より 1 年間 Purdue 大学客員研究員。画像処理, パターン認識・学習, ニューラルネットワーク, 統計的学習, Web データマイニング等の研究に従事。現在, NTT コミュニケーション科学基礎研究所副所長。奈良先端大客員教授。電気通信普及財団賞(テレコムシステム技術賞)(1997, 2006), 本会論文賞(2000, 2004), 情報処理学会山下記念研究賞(2006)等受賞。情報処理学会, 日本神経回路学会, IEEE 各会員。