LD-005

Relative Innovator の発見によるパーソナライズ手法の提案

Personalized Ranking by Identifying Relative Innovators

川前 徳章 Noriaki Kawamae　**山田武士** Takeshi Yamada　**上田修功** Naonori Ueda

1．はじめに

本稿はユーザ間の関係と履歴に基づいて，検索結果のランキングをパーソナライズする手法を提案する．現在，書籍や音楽から，レストランやホテルまで，さまざまな情報を検索できるサービスが Web 上に数多く存在している．これらのサービスでは個々のユーザの興味や目的に適応するために，検索結果のランキングのパーソナライズが重要となる．その手法の一つとして協調フィルタリング(CF)があり，多くの EC サイトで利用されている．CF はユーザの履歴を利用し，過去の行動が類似したユーザ集合は将来も類似した行動をするという考え方に基づく予測手法である．

履歴に基づくパーソナライズでは，あるユーザの将来の行動を予測するのに有用なユーザ集合を見出すため，ユーザ間の関係を適切に定義する必要がある．あるユーザと同じアイテムを先行してアクセスしたユーザの履歴には，そのユーザが将来アクセスすると考えられるアイテムに関する情報が含まれていると考えられる.そこで我々は，このようなユーザを発見するために，ユーザのアクセスの時系列性とアイテムの重要度に基づいてユーザ間の関係を定義する．マーケティング分野において，新製品をユーザ集合の中で他ユーザに先駆けて購買するユーザを Innovator，Innovator と同じアイテムを遅れて購買するユーザを Imitator と呼んでいる[7]．我々はユーザ毎に共通するアイテムを先行してアクセスしたユーザを Relative Innovator と呼ぶ．

ユーザ間で共通する履歴の期間が短い場合にはユーザ間で共通するアイテムが少なくなる場合がある． 従って，ユーザ間の関係を共通するアイテムのみに基づいて定義した場合，所望の Relative Innovator が見出せない可能性がある．この問題を回避するために，ユーザ間の関係をマルコフ連鎖ネットワークとして捉え，ユーザ毎に他のユーザを Relative Innovator として選択する確率として Relative Innovator Degree (RID)を定義する．そして，RID をユーザ間の関係として用い，RID を利用した情報のスコアリングによって検索結果のランキングをパーソナライズする．実データを用いた実験により提案法の有効性を検証する．

2．既存研究

情報検索において，ユーザ間の関係を抽出する為に口コミサイトや SNS 等のユーザ間の時系列的な情報伝播を利用する研究が提案されている． ［7］はユーザ集合において他よりも早くアイテムを購買するユーザを Innovator として特定するノンパラメトリックなアプローチを提案した．［4,8］はネットワーク上での情報伝達を把握するため，購買者間の影響をモデル化している．[9]はユーザ間の情報伝播モデルによって Imitator を発見する方法を提案している．［4］は動的環境におけるユーザの興味を長期および短期の観点から抽出する方法について提案し，［2］はユーザの興味の時間的変化をモデル化するのに忘却係数を用いた．

本稿では，ユーザ間の関係を，情報へのアクセスの時系列性と情報の重要度に着目して定義する．その結果，ユーザ毎に定義するユーザ間の関係は，非対称であって，且つ時系列的に変化し得るという特徴を持つ．従来 Innovator はユーザ全体の中で特にアクセスが早いユーザを指し，既存研究では主に，その抽出自体が目的であり，それらをパーソナライズに利用する試みはなされていない．本稿で提案する Relative Innovator はパーソナライズの精度向上に有効なユーザの選択が目的であり，ユーザ毎に相対的に定義している点が異なる．

3．Relative Innovator の発見によるパーソナライズ

3.1 履歴からの Relative Innovator の発見

ここではパーソナライズにおける Relative Innovator とその履歴の有効性と，その発見に必要な条件をユーザの行動の時系列性と情報の重要度の観点から検討する．

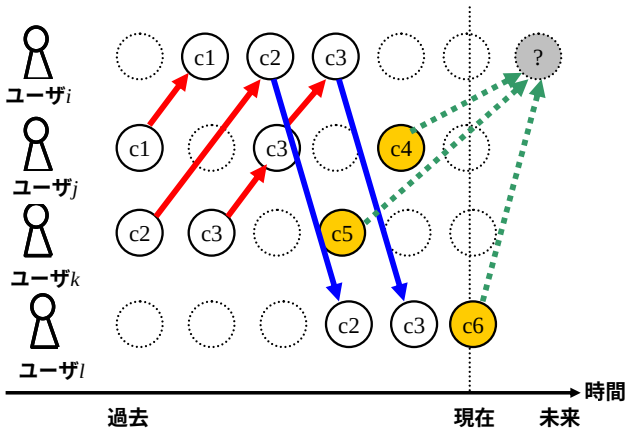

図 1．ユーザのアイテムへのアクセスの時系列性

図 1 はユーザ *i, j ,k, l* の履歴を用いて，各ユーザがアクセスしたアイテムを時系列順に並べたものである．ここで，ユーザ *i* が将来アクセスするアイテムを予測することを考える．つまり，ユーザ *i* が現在までアクセスしていないアイテム c4, c5, c6 のうち，将来どのアイテムを最もアクセスする可能性が高いかを考え，その可能性を数値化してユーザ *i* に対してアイテムをランキングする．

図 1 においてユーザの関係をユーザ行動の時系列性から考える．ユーザ *i* はユーザ *j* 及び *k* と同じアイテムを時系列的に遅れて，ユーザ *l* は更に遅れてアクセスしている．アイテムのアクセス順序を比較すれば，ユーザ *j* 及び *k* はユーザ *i* の Relative Innovator であり，ユーザ *l* はユーザ *i* の Relative Imitator となる．アイテムのアクセスに関して過去にユーザ *j* 及び *k* がユーザ *i* に先行していることから，ユーザ *i* がユーザ *j* 及び *k* と同じアイテム c4 及び c5 にアクセ

スする確率はユーザ *l* と同じアイテム c6 にアクセスする確率よりも高いことが期待される．一方，ユーザ間の関係を従来の CF の代表的な類似度である Pearson を用いて定義すると，ユーザ *i* からユーザ *j, k* そして *l* との関係は全て等しくなってしまう．従って，ユーザ *i* がアイテム c4, c5, c6 へアクセスする確率も等しくなり，上述の期待に反する．ユーザ *l* がアクセスしたアイテムをユーザ *i* がアクセスしていないのは，例えば，既に他のルートでアクセスしたか，知っていても興味がないなど，それなりの理由があると考えられる．従って，ユーザ *i* のランキングにユーザ *l* がアクセスしたアイテムが含まれると，ランキングの精度が低下する恐れがある．従って，パーソナライズの精度向上の為にはユーザの行動の時系列性を考慮する必要がある．

更に図 1 において情報の重要度，ここではアイテムの重要度を考慮することを考える．ユーザ *i* がユーザ *j* 及び *k* とそれぞれ共通してアクセスしたアイテムは二個あるが，アイテム c1 はユーザ *i* と *j* の間のみ，アイテム c2 はユーザ *i* と *k* の間のみでアクセスされている．これらのアイテムの重要度が等しければ，ユーザ *i* から見たユーザ *j* そして *k* への関係も等しいが，一般にアイテムの重要度は異なり，かつ，時間の経過と共に変化する可能性があると考えられる．従って，パーソナライズの為には行動の時系列性に加えて，ユーザ間で共通するアイテムの重要度も考慮する必要がある．

3.2 Direct Relative Innovator Degree

前節で説明した様に，各ユーザから他のユーザを Relative Innovator と見なす確率を定義する際，ユーザの行動の時系列性とユーザ間で共通するアイテムの重要度を同時に考えることが必要である．さらに，図 1 のアイテム c3 にユーザ *k, j, i* の順序でアクセスしており，アイテム c3 において，ユーザ *i* の Relative Innovator はユーザ *j* と *k* であるが，*j* と *k* とではその度合いは異なると考えられる．そこで本稿では，この度合いはアクセス順位の差に比例すると考えることにする．これに従い，アイテム *k* の重要度を w_k，アイテム *k* に対するユーザ *a* とユーザ *i* の正規化したアクセス順位の差を $r_{ai,k}$ として，ユーザ *a* が全ユーザの中からユーザ *i* を Relative Innovator として選択する確率 $P(u_i|u_a)$ を次のように定式化する．

$$p(u_i|u_a) = \frac{\sum_{k \in C_{ai}} w_k r_{ai,k}}{\sum_{i \in U} \sum_{k \in C_{ai}} w_k r_{ai,k}} \qquad (1)$$

ここで C_{ai} はユーザ *a* とユーザ *i* が共通してアクセスしたアイテム集合，*U* は履歴内のユーザ集合である．この式が意味するのはユーザ *i* がユーザ *a* と同じアイテムを，他のユーザより先駆けてアクセスし，かつそのアイテムの重要度が高ければ，ユーザ *i* がユーザ *a* から Relative Innovator と見なされやすくなることである．

式(1)に含まれる，ユーザ *a* とユーザ *i* の正規化したアクセス順位の差 $r_{ai,k}$ は，ユーザ *a* がアイテム *k* にアクセスした順位を $\eta_{a,k}$，ユーザ *i* がアイテム *k* にアクセスした順位を $\eta_{i,k}$ として次のように定式化する．

$$r_{ai,k} = \begin{cases} \dfrac{\eta_{a,k} - \eta_{i,k} + 1}{N_k}, & \text{if } \eta_{a,k} > \eta_{i,k}, \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

ここで N_k はアイテム *k* にアクセスしたユーザの数である．N_k で割ることによって，$r_{ai,k}$ の値は 0 から 1 の間に正規化される．この $r_{ai,k}$ が意味するのはユーザ *a* とユーザ *i* のアクセス順位に差が殆ど無ければ 0 に近くなり，差が大きくなれば 1 に近くなる．この順位は履歴中のアイテム毎に各ユーザのアクセスした時間から決定できる．ユーザの行動の時系列性としてアクセスの順位差に着目することで，3.1 で述べたように各ユーザの Relative Innovator を DRID により選択できると考える．

アイテムの重要度は時間と共に変化すると考えられる．一般に時系列的な減衰関数は時間を変数とした指数関数として定義される[2]．そこで，式(1)のアイテム *k* の重要度 w_k はアイテム *k* がリリース後にパーソナライズされるまでの経過日数を t_k，アイテム *k* がリリースされてからアクセスされるまでの経過日数の平均を T_k として次のように定式化する．

$$w_k = e^{-\frac{t_k}{T_k}} \qquad (3)$$

w_k が意味するのは，アイテムのリリースが古ければ重要度が小さく 0 に近づき，逆に新しければ重要度が高く 1 に近づくということである．ユーザが他ユーザを Relative Innovator と選択する確率は時間の経過と共に変化すると考えられる．場合によっては Relative Innovator と Relative Imitator が入れ替わる可能性もある．そこでこのような問題を回避する為に，本稿では履歴そのものを時系列的に減衰させるのでなく，各アイテムの重要度を時系列的に減衰させることで，最近の傾向を反映することが可能となる．その結果，より的確に Relative Innovator を選択できると考える．

3.3 Relative Innovator Degree

ここでは Relative Innovator の発見の為に DRID を要素とした行列を確率行列にする方法について述べる．DRID はユーザ間で共通するアイテムのみについてユーザ間の関係を定義しているので，直接的に Relative Innovator を発見する指標と考えられる．その場合，共通するアイテムが少ないユーザはそうでないユーザに比較して RID の値が低くなり，Relative Innovator を選択できない恐れがある．例えば，リリース期間の短いアイテムに着目した場合，あるいは履歴の期間が短い場合はユーザ間で共通するアイテムの数は少なくなる．図 1 において履歴の収集期間を変化させると，ユーザ *j,k* はユーザ *l* に対して Relative Innovator として選択されなくなることが分かる．そこでこの問題を回避する為に DRID をベースとした Markov Chain Model を用いることで複数ユーザを介した間接的な Relative Innovator の発見を試みる．その結果，ユーザ *j,k* はユーザ *i* を介してユーザ *l* にとっての Relative Innovator となる．Markov Chain Model

†（社）情報処理学会, IPSJ

‡（社）電子情報通信学会, IEICE

は Social network の分析[5]，Web 構造の解析[3]でも用いられている．

Markov 行列で定義される Markov Chain がエルゴード的であるとき，Markov 行列の定常分布の存在は保証される．Markov 行列がエルゴード的であるためには，行列が確率行列つまり行列の各要素が非負であり，周辺分布の和が 1 となることが必要である．DRID を要素とする Markov 行列は全要素が 0 となる行を含んでいる可能性があるために，この行列は定常分布の存在を保証しない．そこで DRID を要素とする行列を確率行列とするために，従来の Markov Chain Model を用いた手法[9]と同様に，DRID の和が 0 となる列（他のユーザと共通アイテム数が 0 のユーザ）の要素を 1/*N*（N は列の数，ここではユーザの総数）で置き換える．その結果，RID を要素とした Markov 行列 ***P*** は以下の様に Markov 行列 $\overline{P}$ に置き換えられる．

$$\overline{P}_{ij} = \begin{cases} P_{ij} & \text{if } \sum_j P_{ij} > 0, \\ 1/N & \text{otherwise.} \end{cases} \quad (4)$$

次にこの Markov 行列 $\overline{P}$ を変形し，スムージングパラメタ α を持つ既約行列 $\ddot{P}$ として次のように定義する．

$$\ddot{P} = \alpha \overline{P} + (1-\alpha)\frac{ee^T}{N} \quad (5)$$

上の式により，$\ddot{P}$ は確率行列となり，定常分布の存在が保証される．

3.4 Recommendation From Innovator

RIDに複数のユーザを介した間接的な効果を反映するために，前節で定義した $\ddot{P}$ を用いて確率行列$P_{RID}(u_i|u_a)$を次のように定義する．

$$\begin{aligned} & P_{\mathrm{RID}}(u_i|u_a) \\ & = \frac{1}{l}\left(\ddot{P}(u_i|u_a) + \sum_j \ddot{P}(u_i|u_j)\ddot{P}(u_j|u_a) + \sum_{j_{l-1}\cdots j_1} \ddot{P}(u_i|u_{j_{l-1}})\cdots \ddot{P}(u_{j_1}|u_a) \right) \\ & = \frac{1}{l}\left(\ddot{P} + \ddot{P}^2 + \cdots + \ddot{P}^l\right), \end{aligned} \quad (6)$$

ここで(*l*-1)はユーザ間に間接的に介在するユーザ数である．SNS 等のネットワーク上の情報伝播に関する既存研究[8]及び，予備実験の結果から，*l* は 6 程度が適当と考えられる．

次に RID を用いたユーザ *a* に対するパーソナライズは，既存の CF で用いられた予測モデルを一般化した次の式を用いる．

$$p(c|u_a) \propto \sum_i \delta(c|u_i)p(u_i|u_a), \quad (7)$$

ここで δ(*c*|*u*$_i$)はユーザ*i* がアイテム*c*を既にアクセスしていれば 1，そうでなければ 0 となる．従来は式(7)において*P*(*u*$_i$|*u*$_a$)としてユーザ間の類似度が用いられ，各アイテムのスコアはその類似度の総和になっていたの対し，提案手法は*P*(*u*$_i$|*u*$_a$)として RIDを用いた式(6)の確率$P_{RID}(u_i|u_a)$を用いる点が異なる．その結果，ユーザ*a*にとってRelative Innovator と選択されたユーザによってアクセスされたアイテムはスコアが高くなり，このスコアが高い順にアイテムをランク付けすることで，ユーザ*i*にパーソナライズされたアイテムのリストが作成できる．このようにRIDを用いたパーソナライズ手法を我々はRFI(Recommendation From Relative Innovator)と呼ぶ．

4．実験

4.1 実験データ

オンラインの音楽ダウンロード及びビデオ視聴のサービスにおけるユーザの購買履歴を用いて提案手法の有効性を検証した．音楽ダウンロードの履歴は 2005/4/1 から 2006/7/31 までの 683,366 の購買履歴で，その履歴は 84,620 ユーザと 44,527 タイトルを含み，各購買履歴はユーザ ID，購入した曲のタイトルとその ID，アーティスト，購買日時及び価格などから構成される．ビデオ視聴の履歴は 2005/9/1 から 2006/2/28 までの 66,087 の購買履歴で，その履歴は 7,537 ユーザと 4,064 タイトルを含み，各購買履歴はユーザ ID，購買したビデオのタイトルとその ID 及び購買日時などから構成される．

4.2 実験方法と評価方法

ここでのパーソナライズの目的は各ユーザの行動履歴から，各ユーザが購買する音楽あるいはビデオを予測することである．評価実験では，表 1 に示す様に，各実験データを学習データとテストデータに分割した．そして，学習データからユーザ間の RID を計算し，各ユーザに対して RFI を用い，そのスコアに基いてアイテムリストを作成する．そして，通常の top-N precision（適合率）で予測性能を評価した．

ランキングの精度比較のベンチマークとして CF の代表的手法である Pearson と Nearest Neighbor を用いた．さらに提案手法におけるユーザ行動の時系列性(今回の実験ではユーザのアイテムへのアクセス順位)とアイテムの重要度の効果を評価する為に，それぞれにおいてアクセス順位とアイテムの重要度の有無で比較した．

表 1　実験データの構成

	音楽	ビデオ
テストデータ	2006/7/1 以降に一曲でも購入したユーザの 2006/7/1 以降の購買履歴（8,756 ユーザ/6,814 タイトル）	2006/2/15 以降に一曲でも購入したユーザの 2006/2/15 以降の購買履歴（2,579 ユーザ/1,887 タイトル）
学習データ	テストデータに含まれるユーザの 2006/7/1 以前の購買履歴（8,756 ユーザ/24,566 タイトル）	テストデータに含まれるユーザの 2006/7/1 以前の購買履歴（2,579 ユーザ/3,766 タイトル）

4.3 Innovator を用いたランキングによる実験結果

図 2 にビデオ視聴のデータに対する実験結果，図 3 に音楽ダウンロードのデータに対する実験結果を示す．以下の図において Co は共起する数，Pear は Pearson，U0 はアクセス順序，つまり前後関係しか考慮しない場合を意味する．U1 は提案手法であり，(w)はそれぞれの式においてアイテムの重要度を考慮し，そうでない場合はアイテムの重要度を一律に 1 としている．

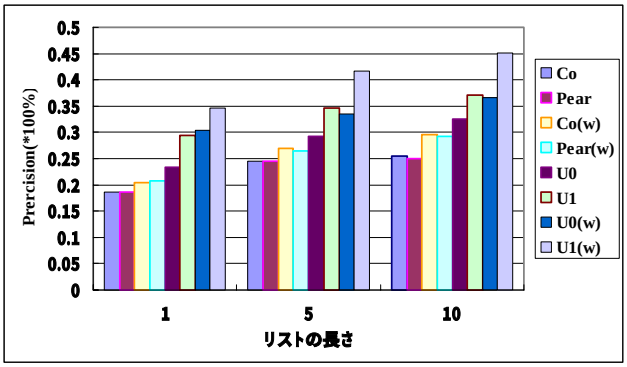

図 2. ビデオ視聴のサービスにおけるパーソナライズの精度比較

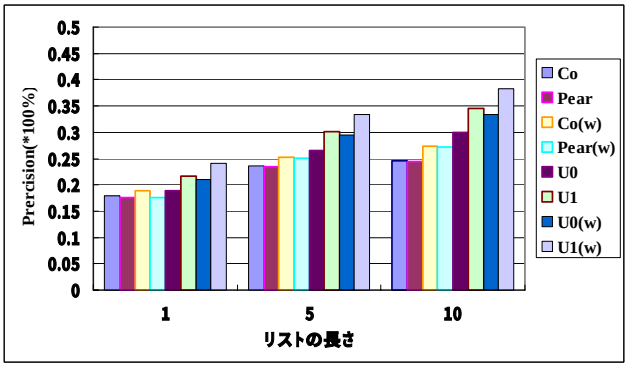

図 3. 音楽ダウンロードのサービスにおけるパーソナライズの精度比較

図 2 に示すように，アイテムの重要度を考慮した場合，考慮しない場合に比べ Co は最大で 16.6%，Pear は最大で 17.0%，U0 で最大で 30.3%そして U1 では 21.6%Precision が向上している．この差はリスト(N)の長さに比例して大きくなっている．この結果より，履歴においてアイテムの重要度を考慮することはパーソナライズの予測精度の向上にとって重要であることが分かる．

次にアイテムの重要度を考慮し，同じ条件の下で，アイテムのアクセス順序を考慮した U1 は，考慮しない Co(w)に対して最大 70.0%，Pear は最大で 67.7%，そして U0 で最大で 24.3%Precision が向上している．こちらでもこの差はリストの長さに比例して大きくなっている．図 3 に示した結果においても同様の傾向が観測された．以上，二つの実験データを通して，パーソナライズの精度向上にはアイテムのアクセス順位とその重要度を考慮した RID が有効であることが確認できた．

5．考察

口コミやSNS等のバーチャルネットワーク上のユーザ間の関係は，ユーザaがユーザiへ確率$P(u_i|u_a)$によるランダムウォークと解釈でき，確率行列からなる遷移行列を構成すると見なすことが出来る．RIDをユーザ間で直接的に定義すると，ユーザa，i間で共通してアクセスしたアイテムが希少な場合，ユーザaがユーザi をRelative Innovatorと見なす確率は 0 に近くなる．この問題を回避する為に，RIDを複数のユーザを介して間接的に定義し，予備実験においてRelative Innovatorの発見における効果を確認できた．

6．まとめ

本稿は情報検索におけるユーザの検索行動を効率化するために Relative Innovator に着目したランキングのパーソナライズ手法を提案した．本稿では Relative Innovator の発見のために RID を定義し，Relative Innovator の履歴を用いてリコメンドを行う方法として RFI を提案した．実験において Relative Innovator を用いたランキングのパーソナライズの有効性を確認した．

参考文献

[1]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, 52(3): 212-225, 2001.

[2]Adomavicius, R. Sankaranarayanan, S. Sen and A.Tuzhilin, Incorporating contextual information in recommender systems using a multidimensional approach, ACM Transactions on Information Systems, 32(1): 103-145, 2005.

[3]M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, 46(5):604-32, 1999.

[4]O'Donovan and B. Smyth, Trust in recommender systems. In Proc. Of the 10th Intl. Conf. on Intelligent User Interfaces, 167-174, 2005.

[5]Scott, Social Network Analysis: A Handbook. Sage Publications, London, 2000.

[6]Domingos and M. Richardson, Mining the Network Value of Customers, In Proc. of the ACM SIGKDD, 2001.

[7]Rusmevichientong, S. Zhu and D. Selinger, Identifying Early Buyers from Purchase Data, In Proc. of the ACM SIGKDD, 2004.

[8]Guha, R. Kumar, P. Raghavan, and A. Tomkins, Propagation of Trust and Distrust, In Proc. of the Intl. World Wide Web Conf., 2004.

[9]Song, C.Y. Lin, B. Tseng, and M.T. Sun: Personalized Recommendation Driven by Information Flow, In Proc. of the ACM SIGIR:509-516, 2006.