3C-4

粒子フィルタを用いた電子商取引サイトユーザの興味分野推定

佐藤 哲 †

楽天株式会社 †

1. はじめに

レコメンデーション，ターゲティング，パーソナライズなど，ネットユーザの嗜好や興味を分析することによるマーケティング手法は多数実用化されており，特に EC サイトではユーザ間の協調フィルタリングが利用されていることが多い．しかし,「嗜好の多様性」と呼ばれる問題により，商品単位の協調フィルタリングではユーザの潜在的な関心を推定することが難しくなっている．本発表では，粒子フィルタを利用し多次元空間でユーザの興味分野を推定・追跡する手法を提案する．粒子フィルタは多次元の計算に速度的に有効なマルコフ連鎖モンテカルロ法を用いており，多次元空間を利用する本研究に適している．本手法は多様な分野にまたがるユーザの興味分野を確率分布の形で保持するため，多様な嗜好性を網羅するマルチカテゴリ問題の一つの解決法として有効である．

2. ユーザの関心分野の推定

ユーザが何に興味を持っているかを推定する手法としては，(1) セマンティック Web のように，タグ等によってジャンル情報が付加されているページの閲覧履歴ログを参照する方法，(2) 協調フィルタリングのように，ジャンル情報の推定が容易なネット購買履歴ログを参照する方法，(3)WordNet や Wikipedia を用いた，検索単語から単語の属するジャンル情報を推定する方法 [1]，等が一般的である．ただしこれらの手法は現在の関心分野の推定が主となることが多く，本研究の目的である未知の関心分野を予測し，ユーザの電子商取引に役立つ情報を提示することに適用することは難しく，また,「嗜好の多様性」に対応するような複数の情報に基づく推定や予測も難しい [2]．本研究の目的に対し適していると思われる，学習や予測を繰り返し，過去から現在の状態を推定しつつ時系列的に予測する手法の一つに，粒子フィルタがある．

粒子フィルタは，確率状態モデルによる予測と観測モデルによる観測データとのデータ同化を繰り返すプラットフォームで，確率分布を多数の粒子で近似する．マルコフ連鎖モンテカルロ法を利用しているところから，モンテカルロ法の特徴である比較的高次元空間での近似計算が抑えられた計算量で可能である [3]．その特徴を利用し，本研究では n 次元多様体上に正規直交座標系を設定し，各座標軸に対し予め用意したジャンルを対応させる．そして原点からの距離を，そのジャンルへの関心の高さと定義する．この設定により，関心分野の移り変わりや複数の分野への関心等を表現する事ができる．本研究では粒子フィルタを用いて，ジャンル情報が付加されている Web ページの閲覧履歴ログを n 次元空間の中で時系列的に追跡することにより，多様な嗜好に対応した関心分野の推定と学習及び予測を行う．

3. 粒子を用いた確率密度近似による関心ジャンルの表現

簡単のため 2 次元空間を想定し，x-軸がジャンル「パソコン・周辺機器」への関心の高さを，y-軸がジャンル「メンズファッション」への関心の高さを表すとし，この平面上でのイメージ図を元に説明する．この場合，ジャンル「メンズファッション」に関心があるユーザに対しては図 1 のような近似確率場が得られ，ジャンル「パソコン・周辺機器」に関心がある場合は図 2 のような近似確率場が得られる．一方，ジャンル「メンズファッション」「パソコン・周辺機器」の両方に関心がある場合は図 3 のような近似確率場が得られ，両ジャンルへの関心の高さを表現できる．ジャンルの数が増えても同様である．

本研究では，粒子フィルタのモデルは主に次のようなシンプルな形式を扱う：

$$\begin{cases} \mathbf{x}_t &= \mathbf{x}_{t-1} + \Delta t \mathbf{v}_t + \nu_t \\ \mathbf{y}_t &= \mathbf{x}_t + \omega_t \end{cases}$$

ここで，$\mathbf{x}_t$ は時刻 t での状態ベクトル，$\mathbf{v}_t$ は時刻 t での状態ベクトルの移動速度，$\mathbf{y}_t$ は時刻 t での観測ベクトル，ν_t は全ての要素が独立した正規分布 $N(0, \sigma_\nu)$ に従うシステムノイズ，ω_t は全ての要素が正規分布 $N(0, \sigma_\omega)$ に従う同一の値を持つ観測ノイズ，α は定数である．つまり，計算コストを抑えるために多次元の正規分布は採用せず，システムノイズは n 個の独立した 1 次元の正規分布に従う要素から成るベクト

Interest genre estimation of EC-site users using particle filter

†Tetsu R. Satoh, Rakuten Inc.

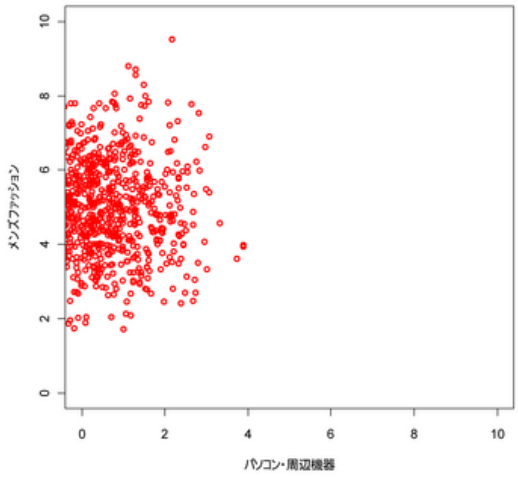

図 1: 例：ジャンル「メンズファッション」に関心がある場合

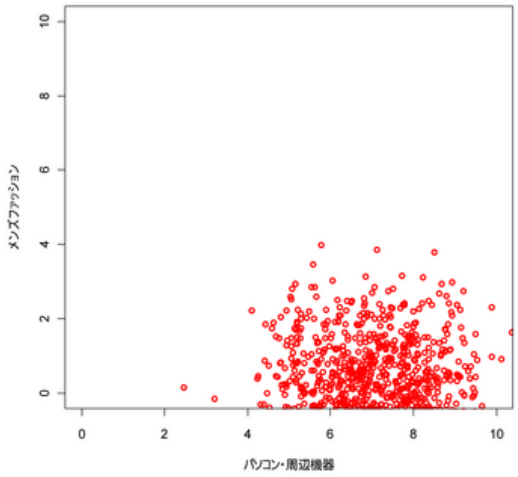

図 2: 例：ジャンル「パソコン・周辺機器」に関心がある場合

ル，観測ノイズは 1 次元の正規分布に従うスカラー値と n 次元の要素が全て 1 のベクトルとの積である．また，Δt が乗じてある項は，関心分野は急激には変化しないという仮定を表しており，関心があると推定されたジャンルの関心度合いをしばらくの間強める働きをする．

関心があるジャンルの測定には，Web ページに商品のジャンル情報が付加されている EC サイトにて，ユーザが閲覧した履歴ログを用いる．ジャンル情報と時間情報のペアを抽出し，時系列的にジャンル情報を座標軸を示す ID 情報に変換したのち，ID 情報の観測積算カウンタ値を観測値として粒子フィルタに入力することで，そのユーザに対する関心のあるジャンルを表す確率分布を作成していく．

実際のデータによる評価結果は，発表当日に示す．

4. おわりに

個別の分野・商品への関心度合いを推定するのではなく，関心があると思われる想定した全てのジャンルへの関心度合いを粒子による近似的な確率分布として保持することにより，ユーザの多様な興味へ対応する興味分野表現・推定手法を提案した．また，確率分布空間として座標軸とジャンルを対応させ，想定ジャンルの数だけ空間の次元を増やすことでマルチカテゴリ問題の解決を図った．多次元空間の導入は粒子フィルタによるデータ同化の手法としては普通に用いられるが，Web 上の EC サイトでのユーザ行動分析にも有効である．本研究ではジャンル分類済みの Web ページの閲覧情報のみを用いたが，よりカバレッジの広い検索キーワードを用いたり，ユーザのネット上の発言や商品等の評価情報へ適用しデータ同化の手法により結果を統合することが今後の課題である．

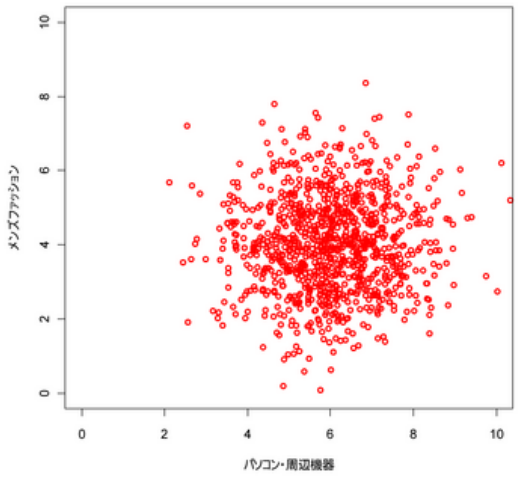

図 3: 例：ジャンル「メンズファッション」と「パソコン・周辺機器」の両方に関心がある場合

参考文献

[1] 佐藤哲．オフライン全文検索エンジンを用いた文字列間の正規化類似度距離．第 10 回情報科学技術フォーラム，Vol. 2, No. F-008, pp. 397–398, 2011.

[2] Sangkeun Lee, Sang-il Song, Minsuk Kahng, Dongjoo Lee, and Sang-goo Lee. Random walk based entity ranking on graph for multidimensional recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, RecSys '11, pp. 93–100, New York, NY, USA, 2011. ACM.

[3] 中村和幸，上野玄太，樋口知之．データ同化：その概念と計算アルゴリズム．統計数理，Vol. 53, No. 2, pp. 211–229, 2005.