6C-1

潜在クラスを利用した クロスメディアレコメンデーション方式の提案

柳原 正† 帆足 啓一郎† 松本 一則† 菅谷 史昭†

†株式会社KDDI研究所

1. はじめに

協調フィルタリングを用いた既存のレコメンデーションシステムでは，履歴情報が無いユーザに対してレコメンド結果が計算できない．この問題を解決するため，履歴情報がないユーザの他履歴情報から潜在クラスモデルを用いて嗜好情報を抽出し，レコメンド結果を計算するクロスメディアレコメンデーション方式を提案する．

2. 従来手法

近年，ユーザに新たな発見を促進するレコメンデーションシステムが頻繁に利用されるようになった．例えば，MovieLens[1]ではユーザが過去に鑑賞した映画の履歴情報を基に，他ユーザの履歴情報と比較し，映画を推奨する．このような推奨方法を協調フィルタリングと呼ぶ．

協調フィルタリングでは，ユーザ間の履歴情報を比較し、相関が高いユーザのアイテムを推奨するという手順でレコメンド結果を作成する．

しかし，この手法では履歴情報が全く無いユーザを対象とした際に，レコメンド結果が計算できない．

3. 提案手法

本研究では，履歴情報が無いユーザに対して高精度なレコメンド結果を生成可能にするため，他アイテムに関する履歴情報からユーザの嗜好情報を抽出し，履歴情報がないアイテムに関する嗜好情報との相関関係を比較する手法であるクロスメディアレコメンデーション方式を提案する．

図 1 にクロスメディアレコメンデーション方式の概念図を示す．本方式では，分析用として用いる履歴情報である分析基データと，推奨したいアイテムに関する履歴情報である推奨基データを用意する．

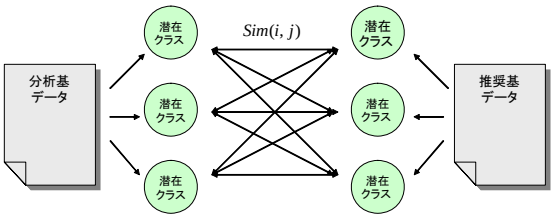

図 1: クロスメディアレコメンデーション方式の概念図

分析基データと推奨基データのそれぞれに潜在クラスモデル[4]を適応することで嗜好情報を抽出する．潜在クラスの抽出方法として EM アルゴリズム[5]を利用する．

潜在クラスが求まれば，それぞれのクラスに含まれるユーザ群を互いに比較し，相関値を求めることで，分析基データの潜在クラスと推奨基データの潜在クラスの相関を算出する．分析基データと推奨基データに共通して出現するユーザ数を n，ユーザ Ui の潜在クラスへの帰属確率を q_t^{Ui} とすると，潜在クラス t は(1)の n 次元ベクトルで表現できる．

$$\vec{t} = (q_t^{U_1}, q_t^{U_2}, ... q_t^{U_n}) \quad \cdots (1)$$

分析基データから抽出された潜在クラス t_a と推奨基データから抽出された潜在クラス t_r はそれぞれベクトルで表現できるため，潜在クラス間の相関値はピアソン相関係数やコサイン類似度のような相関係数を用いることを求める．

それぞれの潜在クラス間の相関値が求まると，推奨基データから抽出された全潜在クラスのうち，レコメンド対象ユーザが帰属する潜在クラスと最も相関が高い潜在クラスを選択し，選択された潜在クラスに帰属する推奨基データの全ユーザのアイテムに関する履歴情報から人気が高いアイテムの上位 k 件を抽出し，出力する．

以上により、従来の協調フィルタリングレコ

"Cross-Media Recommendation Using Latent Class Model"
Tadashi Yanagihara†, Hoashi Keiichiro†, Kazunori Matsumoto†, Fumiaki Sugaya†
† KDDI R&D Laboratories, Inc.

メンド方式において履歴情報が無い場合，レコメンド結果が生成できなかったが，本研究で提案しているクロスメディアレコメンデーション方式ではユーザの嗜好情報を基に推奨を行えるため，より精度が高いレコメンド結果が得られると考えられる．

4. 評価実験

クロスメディアレコメンデーション方式(以下，CM 方式)との性能比較のため，協調フィルタリング方式(以下，CF 方式)を基とした“Item-based Top-N Recommendation Algorithm” [3]を実装したレコメンデーションエンジンを比較対象として用いた．

評価用データとして，GroupLens [2] プロジェクトにて公開されている映画を鑑賞したユーザの 10 万件分の履歴情報を用いた．本実験では最も人気があったジャンルである drama に属するアイテム(725 件)の履歴情報を分析基データとし，drama 以外のジャンルに属するアイテム(657 件)の履歴情報を履歴基データとした．

履歴情報が無いユーザに対するレコメンド結果を評価するため，二つの評価軸を設ける．

一つ目は一度削除した履歴情報の再現性(Match Rate)を設ける．まず，推奨基データに含まれているユーザの 10%をサンプリング用ユーザとして選定し，そのユーザの推奨基データの履歴情報を削除する．次に，レコメンド結果を計算した後，サンプリング用ユーザのレコメンド結果を選定し，削除される前の履歴情報と比較し一致した比率として表現する．

二つ目はレコメンド対象アイテムの有効範囲の広さ(Coverage)である．これは，レコメンド結果に含まれたユニークアイテム数を全アイテム数で割った比率として表現する．

5. 結果と考察

それぞれのレコメンデーションエンジンにおいて，ユーザ一人に対して生成するレコメンデーション結果の件数を k 件 $(10 \le k \le 100)$，分析基データと推奨基データの潜在クラス数をそれぞれ 20 個とし，相関係数はコサイン類似度を用いた．サンプリング用ユーザに関するレコメンド結果を取り出し， Match Rate と Coverage を計測した結果を図 2 に示す．横軸をレコメンド件数とする．

Match Rate において，CM 方式の Match Rate は CF 方式の Match Rate に比べ全体的に上回っており，差分が $k = 100$ のとき，最大 0.203 であった．この結果から，CM 方式が CF 方式よりも履歴情報がないユーザに対し，より高精度なレコメンド結果が計算できたと言える．

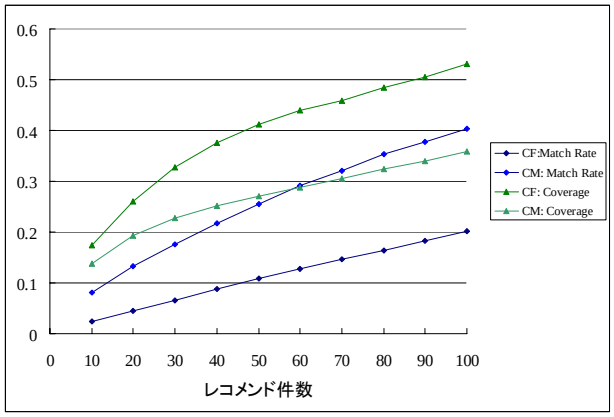

図 2: CF 方式と CM 方式の Match Rate と Coverage

一方， Coverage において，CM 方式の Coverage は CF 方式の Coverage に比べ，全体的に下回っており，差分が $k = 40$ のとき，最大 0.157 であった．CM 方式の Coverage が低かった原因として，分析基データの潜在クラスへの偏りが発生し，推奨されるアイテム数が減少したためと考えられる．

6. まとめ

本研究では，分析基データと推奨基データを基に，潜在クラスを抽出し，潜在クラス間の相関を比較し，レコメンド結果を生成するクロスメディアレコメンデーション方式を提案した．抽出された潜在クラスを基に他潜在クラスと比較し，推奨を行うことで，履歴がないユーザに対し精度が高いレコメンド結果が作成可能であることを示した．しかし，Coverage においては逆に低いことが分かったため，今後は原因と考えられるユーザの帰属するクラスへの偏りを少なくするようにアルゴリズムを改善することを検討する．

7. 参考文献

[1] “MovieLens” http://movielens.umn.edu/
[2] “GroupLens Project” http://www.cs.umn.edu/Research/GroupLens/
[3] M. Deshpande and G. Karypis, “Item-based Top-N Recommendation Algorithms”, ACM Transactions on Information Systems (TOIS) 2004
[4] 岡太彬訓，木島正明，守口 剛, “マーケティングの数理モデル” , 朝倉書店, 2001
[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm.”, J. R. Statistical Society, Series B, 39:1–38, 1977.