D-003

協調フィルタリングによる情報推薦の評価実験を支援する疑似データセットの提案

Simulation data sets to support evaluation experiments of collaborative filtering recommender systems

横田智博 † Tomohiro Yokota
小林亜樹 ‡ Aki Kobayashi

1. 背景と目的

協調フィルタリング (以下 CF) による情報推薦はデータセットの持つ特性を活かして推薦を行う [1] が，公開されている実データセットが少ないため，各提案による情報推薦の特性の違いを知るためには，多くの異なる特性を持ったデータセットが必要である．そこで本研究では，現実に即した特性の異なるデータセットを生成することを目的とする．実データセットから協調フィルタリングに活かせる特性を失わないようにするために，実データセットの類似度を観測し，類似度の分布を再現し得るパラメータを持つ関数を定義し，次にその関数から現実に即したデータセットを生成する，という作成方法を提案する．

2. 疑似データセットの作成方針

本研究では作成する疑似データセットを実データセットの代わりとして情報推薦に用いることができるようにするため，実データセットの中身と推薦時の用いられ方を参考にし，疑似データセットの作成方針を立てる．

まず，実データセットの中身については，ユーザがアイテムへ何件か評価を与えているという状況を示したデータの集まりが必要であり，CF による情報推薦に関する研究で用いられるデータセットとして代表的なものに MovieLensDataSet(以下 MLDS)[3] というものがあり，これも上記の状況を示すデータの集まりになっている．

次に，データセットの用いられ方について，被推薦者とその他のユーザの類似度を算出する際には両ユーザが評価した共通のアイテムの評価値が必要になる．また，推薦したいアイテムの評価値が必要になる [2]．

以上の実データセットの用いられ方を参考にし，本稿ではどのような擬似データセットを作成するのか方針を定めた．

I ユーザが各アイテムへ評価値を与えた状況を示すデータセットを生成する

II 未評価値は生成しない

III 生成するユーザ人数とアイテム数を設定できる

IV 実データのユーザ間の類似度の特性を表現可能である

3. 類似度ヒストグラムを表現する関数

実データのユーザ間の類似度の特性を表現可能とするデータセットを作成するために，関数を導入する．また，本稿では，適当なユーザを推薦ターゲットとした場合の推薦方式の特性の違いや，攻撃に対する耐性などを評価 [4] するプラットフォームとしての利用を目的として，ターゲットとするユーザ毎に異なるデータセットの生成を行うアプローチをとる．MLDS における各ユーザ間の類似度の特性を示したものが図 1 と図 2 である．横軸は類似度，縦軸はユーザ人数の度数を示す．このユーザ 1 と 2 はアイテムへ多く評価を行った上位 200 ユーザ§において，類似度ヒストグラムの中央値が最大値と最小値になったユーザである．

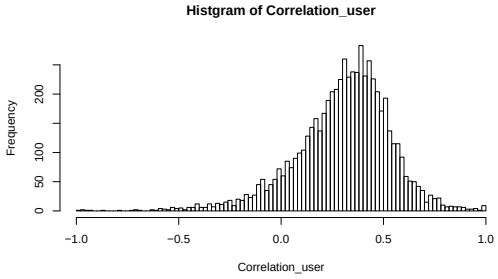

図 1: ユーザ 1 の類似度ヒストグラム

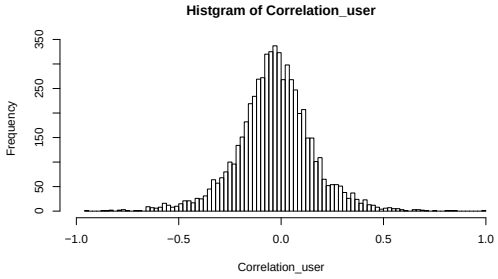

図 2: ユーザ 2 の類似度ヒストグラム

図 1 や図 2 のような各ユーザの類似度ヒストグラムを表現する関数として，ベータ分布の確率密度関数を $\mathrm{beta}(x,p,q)$ としたとき，x の値を変数変換を行い，積分を行ったとき 1 となるように式変形を行った式 (1) の関数を導入し，これを本研究では近似分布関数と呼ぶ．

$$\frac{1}{2}\mathrm{beta}(\frac{x+1}{2},p,q) \tag{1}$$

図 3 は実線でプロットした分布が図 1 を，点線でプロットした分布が図 2 を近似分布関数によって表現したものである．図 3 の実線でプロットした分布のパラメータは

$$p=11,\ q=5.6 \tag{2}$$

であり，点線でプロットした分布のパラメータは

$$p=11.2,\ q=12 \tag{3}$$

であると実験により得た．

式 (2) と式 (3) より，p と q の関係式を

$$p=32q-372.8 \tag{4}$$

†工学院大学 大学院工学研究科 電気・電子工学専攻
‡工学院大学 工学部 情報通信工学科

§ユーザ 1 は 733 件，ユーザ 2 は 849 件分各アイテムへ評価を行っており，それぞれ 150 位と 86 位に位置する．

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.

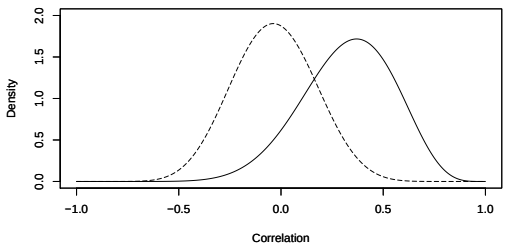

図 3: 近似分布関数による分布

とみなし，q の値のみで類似度ヒストグラムの表現が可能とする．すなわち，類似度ヒストグラムのピークを最も高い類似度となるようなユーザをシミュレートする際には $q = 5.6$ を，逆に最も低いユーザとする場合には $q = 12$ とし，これらの中間の値を自由に設定できる．

4. 疑似データセット生成手順

全体の流れとして STEP1．近似分布関数パラメータを設定し分布の作成を行う．STEP2．近似分布関数の分布に基づいて各階級ごとに評価値列を生成する．という手順をとる．各手順で用いる文字式と定義を示し，図 4 に疑似データセット生成手順のフローチャートを示す．

q：近似分布関数のパラメータ
r：近似分布関数による確率分布の総階級数
I_{total}：アイテム総数
U_{total}：生成するユーザ総数
L：階級を示すカウンタ
U_t：評価値列生成時もとになるユーザ
tXi：U_t の i 番目の評価値
$\bar{U}_t$：U_t との相関が-1 になる評価値列を持つユーザ
$\bar{t}Xi$：$\bar{U}_t$ の i 番目の評価値
U_s：U_t をもとにして生成されたユーザ
sXi：U_s の i 番目の評価値
C：比較基準になる相関値を示す値
C'：評価値列間で相関値を計算した値

STEP 1. の手順

STEP 1.1 q，r，I_{total}，U_{total} の値を設定する

STEP 1.2 式 (1) に q を代入し p を算出する

STEP 1.3 p と q の値を式 (3) に代入し近似分布関数の分布を決定する

STEP 2. の手順

STEP 2.1 $i = 1$，$C' = 1$，$C = 1$，$L = 1$ とする

STEP 2.2 tX_i を I_{total} 個生成する

STEP 2.3 $\bar{t}X_i$ を I_{total} 個生成する

STEP 2.4 sX_i を I_{total} 生成する

STEP 2.5.1 $C' \leq C$ の判断を行い，偽なら 2.5.2 へ真なら 2.6 へ進む

STEP 2.5.2 sX_i を $\bar{t}X_i$ に置き換える

STEP 2.5.3 U_t と U_s の C' を算出する

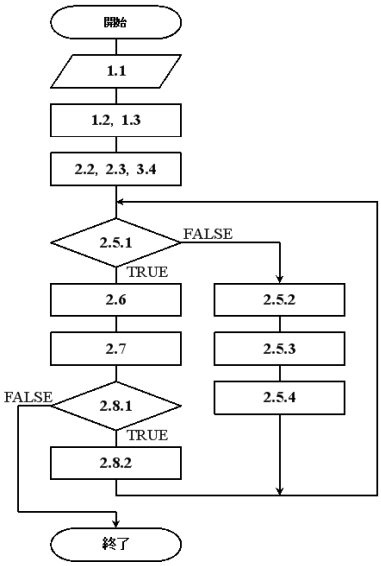

図 4: 疑似データセット生成フローチャート

STEP 2.5.4 $i = i + 1$ とし 2.5.1 に戻る

STEP 2.6 U_s を $\frac{1}{r}U_{total}\mathrm{beta}(\frac{C+1}{2}, p, q)$ 個生成する

STEP 2.7 $L = L + 1$，$x = x - \frac{2}{r}$ とする

STEP 2.8.1 $L \leq r$ の判断を行い真なら 2.8.2 へ，偽なら 2.9 へ進む

STEP 2.8.2 U_s をもう 1 つ生成し 2.5.1 に戻る

STEP 2.9 処理終了

5. まとめと今後の課題

本稿では実データセットにおける各ユーザの類似度の特性を表現可能な疑似データセットの生成方法の提案を行った．また，類似度ヒストグラムのピークを表現し得るパラメータは $5.6 \leq q \leq 12$ であると実験により明らかにした．

今後の課題として，類似度ヒストグラムの尖度を表現し得る関数やパラメータの推定，疑似データセットを用いて情報推薦を行った際の推薦結果の変化の調査などが挙げられる．

参考文献

[1] Herlocker, Jonathan L. Konstan, Joseph A. and Terveen, Loren G. and Riedl, John T., “Evaluating collaborative filtering recommender systems”, ACM TOIS, Vol. 22, pp. 5-53, (2004).

[2] Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl “Item-based Collaborative Filtering Recommendation Algorithms”, Proc. 10th Int. Conf. WWW, pp. 285-295, (2001).

[3] GroupLens Research , MovieLens Data Sets , `http://www.grouplens.org`

[4] S. Lam and J. Riedl,“ Shilling Recommender Systems for Fun and Profit, ”Proc. 13th Int. Conf. WWW, ACM Press, pp. 393-402, (2004).

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.