情報推薦における近傍ユーザの自動選択

情報推薦における近傍ユーザの自動選択

情報推薦における近傍ユーザの自動選択

○三浦功輝　武内充　岡田吉史（室蘭工業大学）

概要　近年、ユーザの意図や嗜好に合った情報を提示する情報推薦技術が盛んに研究されている.我々は以前,漸進的変数選択法を用いて推薦要求ユーザと嗜好の似たユーザ(近傍ユーザ)を特定する手法を開発した.本研究では,新たに近傍ユーザと非近傍ユーザを自動的に決定する手法を導入した.本発表では,本手法を導入した推薦精度の結果について報告する.

キーワード: 情報推薦, 変数選択法

1. はじめに

近年, ユーザの意図や嗜好に合った情報を提示する情報推薦技術が盛んに研究されている. 中でも, 協調フィルタリング法は, 他人の評価に基づく推薦手法であり, 実装が容易なため, 多くのオンラインショッピングサイト等で利用されている. 協調フィルタリングの1つであるユーザベース方式は, 推薦を受けるユーザ（推薦要求ユーザ）と好みの似ているユーザ（近傍ユーザ）を選出し, それらのユーザが高く評価した商品や情報（以下, アイテム）を推薦する方法である. 近傍ユーザを選出する処理は, 近傍形成と呼ばれる. ユーザベース方式では, 推薦要求ユーザと近傍ユーザとの間の好みの類似度は, 推薦要求ユーザが評価した全てのアイテムに対する評価値を用いて計算される. すなわち, これは, 推薦要求ユーザが評価した全てのアイテムに渡って評価が似ている近傍ユーザが存在する, という仮定に基づいている. しかし一般にアイテムの個数は非常に多いため, 現実問題として上記の仮定が常になりたつとは言い難い.

以前, 我々は, 一部のアイテム集合に対する評価が極めて類似しているならば, そのようなユーザは嗜好も類似していると仮定し, これに基づいて近傍形成を行う手法を提案した. 以前の手法では, 近傍ユーザと非近傍ユーザを識別する際の類似度の閾値として, 固定値を使用していた. しかしながら, 推薦要求ユーザごとに, その他のユーザとの類似度の分布は異なるため, 適切に近傍ユーザの選出が行われないという問題が存在した. そこで, 本研究では, 近傍ユーザと非近傍ユーザを識別する類似度の閾値をユーザの類似度の分布に応じて自動的に決定する方法を導入する. 実験として, 映画推薦データへの適用を行い, 本手法による近傍形成の有効性を考察する.

2．方法

2.1 データセット

本研究では, 情報推奨用のベンチマークデータセットである MovieLens[1] および Jester[1] を用いる. MovieLens には, 900 ユーザの 1682 の映画に対する5段階評価データが格納されている. 各ユーザは最低 10 件の映画を評価しており, 評価数は合計 10 万件である. Jester には 3000 ユーザの 100 のジョークに対して-10. 0～10. 0 の範囲の評価データが格納されている. 各ユーザはすべてのジョークを評価しており, 評価数は合計 30 万件である. なお, 以下本稿では, 紙面の都合により, MovieLens を用いた場合について記述する.

2.2 漸進的変数選択法を用いた情報推薦における近傍形成

Fig. 1 に本研究の手法の概要を, 以下に手順を示す.

(1)近傍, 非近傍の分類

各ユーザの各アイテムに対する評価値で構成されるベクトルを評価値ベクトルと呼ぶ. まず, 推薦要求ユーザと他の全てユーザとの間の類似度を評価値ベクトルを用いて計算する. 類似度として, コサイン尺度を使用する. 推薦要求ユーザAuとその他のユーザU_jとのコサイン尺度は１）式により算出される.

$$\mathrm{sim}(\mathrm{Au}, U_j) = \frac{\overrightarrow{Au} \cdot \overrightarrow{U_j}}{|\overrightarrow{Au}| \times |\overrightarrow{U_j}|} \qquad \cdots 1)$$

続いて, 類似度にしたがってその他のユーザを降順で並べ替え, 類似度が閾値以上のユーザを近傍ユーザと設定し, 閾値未満のユーザを非近傍ユーザと設定する. 類似度の閾値は, 推薦要求ユーザの類似度

Fig. 1 :Outlines of our method.

の分布から算出した中央値，または$1/\sqrt{2}$を比較して値が高いものを閾値と設定する．

(2)アイテム選択

アイテムのそれぞれに対し，k番目のアイテムに関する近傍ユーザと非近傍ユーザの間の相関比$\eta_k{}^2$を下式により計算する．

$$\eta_k{}^2 = \frac{Sb}{St} \qquad \cdots 2)$$

ここで，Stは全変動を表し，Sbは級間変動を表す．全変動は３）式，級間変動は４）式で計算する．

$$\mathrm{St} = \sum_j \left(x_{jk} - \overline{x(k)}\right)^2 \qquad \cdots 3)$$

$$Sb = \sum_c \left(\left(\overline{x_c(k)} - \overline{x(k)}\right)^2 \cdot n_c(k)\right) \cdots 4)$$

ここで，c=1, 2 であり，c = 1 のときは近傍ユーザの集合を表し，c = 2 のときは非近傍の集合を表す．x_{jk}は j 番目のユーザの k 番目のアイテムの評価値を表す．$\overline{x_c(k)}$は k 番目のアイテムにおける近傍ユーザと非近傍ユーザの評価値の平均値を，$\overline{x(k)}$はk番目のアイテム全体の評価値の平均値を表す．$n_c(k)$は k 番目のアイテムにおける近傍ユーザと非近傍ユーザが評価しているユーザ数を表す．

次に，近傍ユーザと非近傍ユーザの間の差を表す統計量である F0 値を下式により求める [2]．

$$F0 = \frac{(n-p)\eta_k{}^2}{(p-1)(1-\eta_k{}^2)} \qquad \cdots 5)$$

ここで，n はアイテムの評価をしているユーザの人数，p はユーザ群の種類数を表す．上式で与えられる F0 値が 2 以上のとき，近傍ユーザ群と非近傍ユーザ群の間に統計的な有意差がある（有意水準 5%）と考えることができる [3] [4]．そこで，F0 値が 2 以上のアイテムのみを残し，それ以外は除外する．

(1)の処理と(2)の処理を繰り返し，終了条件を満たすならば最終的に残ったアイテム群と，これを用いて選出された近傍ユーザ群を出力する．本研究での終了条件は，繰り返し回数が 30 回を超えた場合または，(2)の処理にてアイテムが得られなかった場合とする．

３．評価実験

毎回の更新で得られるアイテム群および近傍ユーザ群を用いた推薦精度の変化を観察することにより，本手法の有効性を考察する．実験は，MovieLens のデータセットを用いた 3-fold cross-validation によって実施される．各更新時における推薦精度の算出は以下の手順で行われる．まず，評価値データベースをユーザに関して3分割し，そのうちの１つをテストデータセット，残り 2 つを訓練データセットとする．次に，テストデータセットから１人ずつユーザを取り出して推薦要求ユーザとする．次に推薦要求ユーザが評価しているアイテムから，アイテム更新を行うための初期アイテム（クエリアイテム）をランダムに選択し，残りのアイテムを推薦精度を算出するためのアイテム群とする．クエリアイテムのみを使用し，前章の方法に従って訓練データセットを用いて近傍ユーザを特定する．続いて，下式によりアイテム k の推薦スコアを算出し，この値が高い順に上位 100 個のアイテムを推薦する．

$$\mathrm{Score}(k) = \overline{Au} + \frac{\sum_{U_j \in z}\left(\mathrm{sim}(\mathrm{Au}, U_j)\cdot\left(u_{jk} - \bar{U}_j\right)\right)}{\sum_{U_j \in z}\left|\mathrm{sim}(\mathrm{Au}, U_j)\right|} \qquad \cdots 6)$$

ここで，zは近傍ユーザのうち，アイテムkに対して評価をしているユーザの集合を表す．$\overline{Au}$は推薦要求

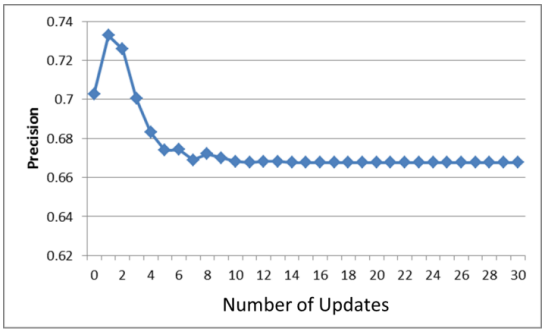

Fig. 2 : Averaged Precision in each update.

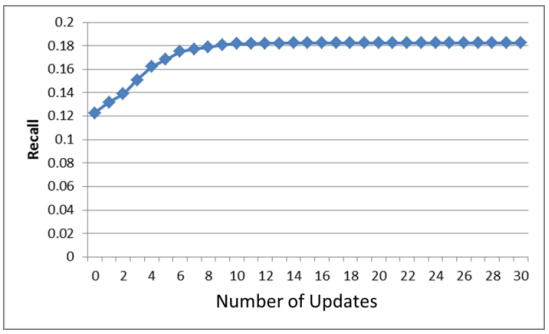

Fig. 3 : Averaged recall in each update.

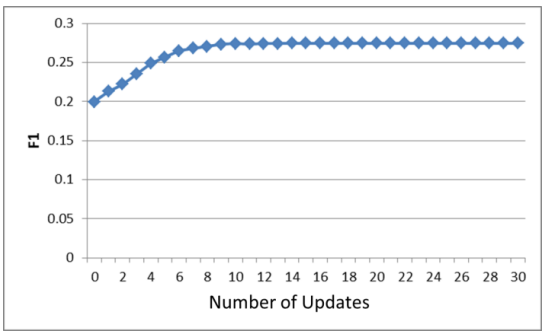

Fig. 4 : Averaged F1 measure in each update.

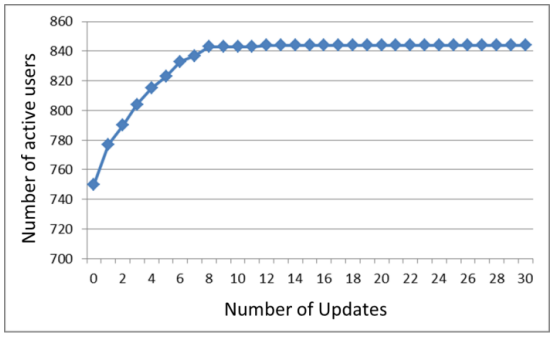

Fig. 5 :Number of active users who are recommended highly-rated items.

ユーザの評価値の平均を，$\bar{U}_j$はｊ番目の近傍ユーザの評価値の平均を表す．u_{jk}は近傍ユーザU_jのアイテムｋに対する評価値を表す．

最後に，アイテム推薦の精度を評価する．本実験では，適合率（Precision），再現率（Recall），F1 値を用いて精度を見積もる[5]．適合率は推薦されたアイテムのうち実際に好きなアイテムの割合，再現率は実際に好きなアイテムをどれだけ多く推薦できたかを示す割合である．F1 値は適合率と再現率の両方を加味した指標である．いずれの値も 0 から 1 の間をとり，値が大きいほど推薦精度が高いことを意味している．

以上の操作をテストデータセットと訓練データセットを入れ替えることで3回繰り返す．最終的な推薦精度は 3 回の試行で得られた値を平均することで算出される．

４．実験結果および考察

Fig. 2 に本法を用いて算出した適合率，Fig. 3 に再現率，Fig. 4 に F1 値を示す．ここで横軸は更新回数を表し，縦軸は推薦精度（適合率と再現率，F1 値）である．この図において，アイテム更新回数=0 とは，全てのクエリアイテムを用いて類似度計算を行う従来のユーザベース方式を意味する．この図から，適合率に関しては，1 回目の更新で 3%程度の上昇を見せ，その後 13 回目の更新までで 8%程度低下し，14 回目以降は平坦に推移している．再現率に関しては，13 回目の更新までに 6%程度上昇し，14 回目以降は平坦に推移している．また，適合率と再現率の両方を考慮した F1 値も再現率と同様な推移を見せていることから，本手法が推薦精度の向上に有効に機能しているといえる．

次に，Fig. 5 に更新回数毎に高評価のアイテムが 1 つ以上推薦された推薦要求ユーザの人数を示す．この図から，11 回目の更新までは，ユーザ数が単調に増加していることがわかる．これは，アイテムの更新を行うことで，より多くのユーザに対し，嗜好に合ったアイテムの推薦が可能になることを意味している．例えば，更新回数 0 回目と 11 回目を比較すると約 100 名の開きがある．本研究で使用したデータセットの全ユーザ数は 900 名であることを考えると，本手法によるアイテム選択の効果は大きいといえる．

このことから，アイテムの更新を行うことで，より多くのユーザに対し，嗜好に合ったアイテムの推薦を行うことができているといえる．今後は，異なるデータセットに対して，本手法を適用し，有効性を検討する必要がある．

５．まとめ

本研究では，推薦要求ユーザとその他のユーザとの間の類似度の分布から，近傍ユーザを選出する閾値を自動的に決定する手法を提案し，アイテムの更新回数と推薦精度（適合率，再現率，F1 値）の関係を調査した．結果，従来のユーザベース方式との比較をとおして以下のことが分かった．適合率は最初の更新で急激に増加，その後減少していき，最終的に収束した．再現率と F1 値はアイテムの更新とともに単調増加し，その後収束した．また，本手法を用いることで，より多くのユーザに対して，嗜好に合ったアイテムを推薦できることが分かった．

今後は，MovieLens 以外のデータセットの適用を行い，本法の有用性を確認していく．

６．参考文献

1) GroupLens_Research: http://www.grouplens.org
2) 判別分析例題: http://ifs.nog.cc/gucchi24.hp.infoseek.co.jp/HANBETUEX.htm
3) 三林光：マイクロアレイを用いた病理診断に有効な遺伝子抽出手法に関する研究，博士論文，室蘭工業大学，(2008).
4) Xinping Wang, Tomomasa Nagashima, Kentarou Fukuta, Yoshifumi Okada, Masahiro Sawai, Hidenori Tanaka And Takashi Uozumi: “Statistical method for classifying cries of baby based on pattern recognition of power spectrum”, InternationalJournal of Biometrics, **2**, 2, 113/123,(2010).
5) D.Jannach, M.Zanker, A.Felfernig, G.Friedrich 著，田中克己，角谷和俊　監訳：情報推薦システム入門理論と実践，p187,p188，(2012).