The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10 - 13, 2002

観測値の類似度を考慮した協調フィルタリング

On a Collaborative Filtering Method using Improved Similarity of Observed Value

†大島 敬志 Keishi Ohshima
‡鈴木 誠 Makoto Suzuki
†平澤茂一 Shigeichi Hirasawa

Abstract—Recommender systems apply knowledge discovery techniques to the problem of making personalized recommendations for information, products or services during a live interaction.These systems especially collaborative filtering ones.are achieving widespread success.Conventional collaborative filtering systems use methods that are based solely on obserbed scores for items.

In this paper we will propose a new collaborative filtering method using improved similarity of value that are based on obserbed binary scores for items.

Keywords—Collaborative Filtering, Recommender Systems,Memory Base Reasoning

1 はじめに

近年，膨大なデータから利用者や消費者等のユーザの要求を自動的に推定し，その要求を満たす情報を積極的に推薦するシステムが研究されている．例えば流通業では書籍やCDについて，過去のユーザの評価データからユーザの要求を満たす商品を推薦している．本論文ではこのような推薦システムの基本技術である協調フィルタリング [4] について考察する．

協調フィルタリングとは過去の購買履歴のデータから購買やアクセスパターンの類似するユーザを同定し，それらのユーザの嗜好から特定ユーザの嗜好を推定する手法である．協調フィルタリングの手法としてよく知られたアルゴリズムに相関係数法 [1] や二項関係学習法 [2] がある．これらの手法ではそのアイテム（商品）が好きか嫌いかという意味での多段階の「嗜好についての類似度」を用いて特定ユーザの嗜好を予測している．しかしアイテムが商品や映画の場合の推薦システムでは「嗜好についての類似度」とは別に，どのアイテムについて回答しているかという意味での1，0で表される「回答パターンの類似度」も存在している．また従来手法では「嗜好についての類似度」を求める際に全てのアイテムを同等に扱っている．しかしアイテムによっては嗜好に影響を与えないものや一部のアイテムには嗜好に強い影響を与えるものが存在している．

そこで本論文では相関係数法と二項関係学習法に対し，「回答パターンの類似度」を用いて嗜好に影響を与えているアイテムに対してウエイトづけを行い推薦システムの精度が向上する事を示す．

2 協調フィルタリング

2．1 協調フィルタリングとは [4]

協調フィルタリングとは，あるアイテムに対するユーザの評価を，そのユーザの別のアイテムに対する評価値と，他ユーザの評価データに基づいて予測することである．具体的な適用例として，映画（アイテム）に対する多数のユーザの評価を入力とし，あるユーザがまだ見ていない映画に対する評価を，そのユーザの他の映画に対する評価と，他のユーザの評価に基づいて予測する．この時入力された評価の数はデータベース全体から見ると非常に少ない．協調フィルタリングの考え方は，評価対象ユーザとの類似度が高いユーザが高評価を与えたアイテムは，評価対象ユーザも高い評価を付ける可能性が高いというものである．

図1に協調フィルタリングの概念図を示す．図1では，評価対象ユーザu_iのアイテムv_xに対する5段階評価値を予測している様子が示されている．ここで未観測の要素は空白で示す．図1に示されたu_i以外のユーザu_1，u_2，u_3のうちu_iと最も評価の傾向が類似しているユーザはu_2であり，u_2のv_xに対する評価が高いことから，u_iのv_xに対する評価も高いものと予測される．またu_1，u_3はいずれもu_iとは評価値の傾向が異なっており，u_iとの類似度が低いことから，v_xの評価値予測の際にはそれほど考慮されない．

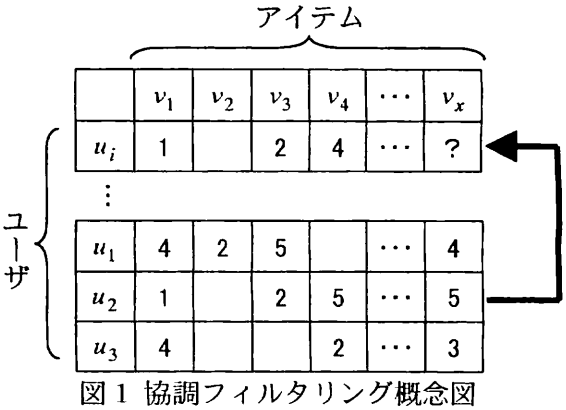

図1 協調フィルタリング概念図

†〒169-8555 東京都新宿区大久保 3-4-1 早稲田大学大学院理工学研究科 経営システム工学専攻．Dept. of Industrial and Management System Engineering School of Science and Engineering, Waseda University, Okubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 JAPAN.
‡〒251-8511 神奈川県藤沢市辻堂西海岸 1-1-25 1410-3 湘南工科大学 情報工学科．Shonan Institute of Technology, Tujido Nisikaigan 1-1-25 Fujisawa-shi Kanagawa-ken, 251-8511 JAPAN.

2．2　従来手法

2．2．1　相関係数法［1］

行列$M=(M_{iz})$は行がユーザを列がアイテムを表す行列として，その(i,z)要素がi番目のユーザのz番目のアイテムの評価値とする．未観測の要素（空白）については欠損値として扱い，$M_{iz}=*$とする．通常，行列Mは欠損値を多く含み，ほとんどのM_{iz}には値が与えられない．相関係数法では評価対象ユーザiのz以外のアイテムに対する評価値と，他のユーザの評価値に基づいて未知の評価値M_{iz}を以下のように算出する．

$$M_{iz}=M_i+\frac{\sum_j C_{ij}(M_{jz}-M_j)}{\sum_j|C_{ij}|} \quad (式1)$$

ここでM_iとM_jはユーザiとjの評価値の平均値

$$M_i=\frac{\sum_x M_{ix}}{\sum_x 1} \quad (式2),\quad M_j=\frac{\sum_x M_{jx}}{\sum_x 1} \quad (式3)$$

である。ここでの$\sum_x$は評価値が既知のアイテムだけについてとる．またC_{ij}は評価対象ユーザiとデータベース中のユーザjの相関係数

$$C_{ij}=\frac{\sum_x(M_{ix}-M_i)(M_{jx}-M_j)}{\sqrt{\sum_x(M_{ix}-M_i)^2\sum_x(M_{jx}-M_j)^2}} \quad (式4)$$

である．ここでの$\sum_x$は評価対象ユーザiとデータベース中のユーザjの両方で評価値が既知のアイテムについてとる．相関係数法にはその他，ユーザの類似度ではなくアイテムの類似度を用いる手法［3］等も研究されている．

2．2．2　二項関係学習法［2］

今，w_{ij}を行iからjへの重みとする．評価値が入力されている要素を参照し，重みw_{ij}を以下のように更新する．

weight update: $M=(M_{jz})$が欠損値でない全ての$j(\neq i)$について

$$\begin{cases} w_{ij}=(2-\gamma)w_{ij} & if\ M_{jx}=M_{ix} \\ w_{ij}=\gamma w_{ij} & if\ M_{jx}\neq M_{ix} \end{cases} \quad (式5)$$

ここでγはユーザが任意に与えた値（ただし$0<\gamma\leqq 1$）である．

観測データを正しく予測するのに寄与している行の重みを$2-\gamma$倍し，誤って予測するのに寄与している行の重みをγ倍する．つまりw_{ij}は以下の式と同義である．

$$w_{ij}=(2-\gamma)^a\gamma^b \quad (式6)$$

ただしaはユーザiに対してjが同じ評価を示した回数，bはユーザiに対してjが異なった評価を示した回数である．実際に予測を行う場合はw_{ij}の和が最大になる値を予測値とする．

$$M_{iz}=\arg\max_{a\in A}\sum_{j:M_{jz}=a} w_{ij} \quad (式7)$$

ただしAは評価値の集合である．

2．3　従来手法の問題点

従来手法では類似度を計算する際に全てのアイテムを同等に扱って類似度を計算している．しかしアイテムによっては類似度に影響を与えるアイテムとそうでないアイテムが存在している．例えば，ホラー映画の評価値の予測を行いたいのに恋愛映画の評価値を用いて類似度の予測を行っても精度は向上しない．その問題点を改善するにはアイテムのコンテンツが分かっていればいいが，コンテンツの分類にはさまざまな概念からさまざまな分類ができてしまい，一意に正解・不正解が定まらない．またコンテンツによるフィルタリングを導入すると，人の嗜好を考慮するのではなく同じコンテンツのアイテムに同じ評価値が予測される傾向が強くなり問題である．

一方，従来手法ではそのアイテムが好きか嫌いかという意味での多段階の「嗜好についての類似度」を用いて特定ユーザの嗜好を予測している．しかしアイテムが商品や映画の場合の推薦システムでは「嗜好についての類似度」とは別に，どのアイテムについて回答しているかという意味での1，0で表される「回答パターンの類似度」も存在している．

次節ではこの問題点を踏まえ，「回答パターンの類似度」を用いてアイテムごとにウエイト付けを行い，予測精度の向上を目指す．

3　提案手法

3．1　提案手法の概要

提案手法ではアイテムzの嗜好に大きな影響を及ぼすアイテムxとは，アイテムzに評価値を入力しているユーザがアイテムxに評価値を入力しており，逆にアイテムzに評価値を入力していないユーザが評価値を入力していないアイテムだと考える．なぜなら同じような傾向を持っているアイテムであればユーザがそのアイテムに興味を持っていれば両方に回答し，逆に興味を持っていなければ両方に回答しないからである．このアイテム間の影響を考えたアイテムの「回答パターンの類似度」を用いて2つの手法を提案する．

①相関係数法に用いることでアイテムzの嗜好に強く影響を与えるアイテムを考慮した「嗜好についての類似度」を求める．

②二項関係学習法に用いることでアイテムzに対してユーザのウエイトを更新する際にアイテムzの嗜好に強く影響を与えているアイテムほどウエイトを大きく更新する．

提案手法のアルゴリズムの概要は以下の通りである．

ステップ1：アイテムについて回答しているかという意味での1，0で求められる評価値を用いてアイテムの「回答パターンの類似度」を求める．

ステップ2：アイテムの「回答パターンの類似度」を考慮し，①では「嗜好についての類似度」を，②で

はウエイトを求める．

ステップ3：①では「嗜好についての類似度」を，②ではウエイトを用いて未知の評価値 M_{iz} を予測する．

3．2 提案手法①（相関係数法）

従来手法ではユーザの相関係数 C_{ij} を求める際に全てのアイテムを同等に扱っている．そこで前節で述べたアイテムの「回答パターンの類似度」を用いることでより正確な「嗜好についての類似度」を求める．

（ステップ 1）行列 $M'=(M'_{iz})$ は行がユーザを表し，列がアイテムを表す行列として，その (i,z) 要素が i 番目のユーザが z 番目のアイテムに評価値が入力されていれば1，そうでなければ0を入力する．

$$M'_{iz}=\begin{cases}0 & if\ M_{iz}=* \\ 1 & if\ M_{iz}\geq 1\end{cases} \quad (式8)$$

行列 M' の値を用いてアイテム間の相関を求める．ここでアイテム間の類似度は内積を用いる．

$$d_{xz}=\frac{\sum_j M'_{jx}\cdot M'_{jz}}{\sum_j M'_{jx}\sum_j M'_{jz}} \quad (式9)$$

d_{xz} はアイテム z に評価値を入力しているユーザが評価値を入力しているアイテムほど値が大きくなる．d_{xz} はアイテムの「回答パターンの類似度」でありこの値が大きいほどアイテム z の嗜好に影響を与えていると考えられる．

（ステップ 2）d_{xz} の値を用いてアイテム z の予測を行う際の「嗜好についての類似度」C'_{ijz} を以下のように修正する．

$$C'_{ijz}=\frac{\sum_x d_{xz}(M_{ix}-M_i)(M_{jx}-M_j)}{\sqrt{\sum_x d_{xz}(M_{ix}-M_i)^2\sum_x d_{xz}(M_{jx}-M_j)^2}} \quad (式10)$$

（ステップ 3）この C'_{ijz} を用いて未知の評価値 M_{iz} を以下のように算出する．

$$M_{iz}=M_i+\frac{\sum_j C'_{ijz}(M_{jz}-M_j)}{\sum_j|C'_{ijz}|} \quad (式11)$$

3．3 提案手法②（二項関係学習法）

従来手法では重みを学習する際に γ というユーザが任意に与える値を用いていた．しかし本論文では γ のかわりに（式9）で求めたアイテムの類似度を用いる．

（ステップ 1）相関係数法と同じく（式9）から d_{xz} を求める．

（ステップ 2）重み w'_{ijz} を以下のように更新する．

weight update: $M=(M_{jz})$ が欠損値でない全ての $j(\neq i)$ についてアイテム z の予測を行う際のウエイト w'_{ijz} について

$$\begin{cases} w'_{ijz}=\{2-(1-d_{xz})\}w_{ij} & if\ M_{jx}=M_{ix} \\ w'_{ijz}=(1-d_{xz})w_{ij} & if\ M_{jx}\neq M_{ix}\end{cases} \quad (式12)$$

とする．

（ステップ 3）実際に予測を行う場合は w'_{ijz} の和が最大になる値を予測値とする．

$$M_{iz}=\arg\max_{a\in A}\sum_{j:M_{jz}=a} w'_{ijz} \quad (式13)$$

4 提案手法の評価と考察

4．1 評価データ

評価データには協調フィルタリングの検証用データとしてよく用いられる MovieLens [5]データとEachMovie[6]データを用いた．

（1）MovieLens データは 943 人の 1682 の映画に対する5段階評価データである．評価されている項目は全部で 100000 個あり欠損率は 93.7%である．ここではそのうち 80000 個をランダムに取り出し学習データとみなし，残りの 20000 個のデータで推定誤差を評価した．

（2）EachMovie データは 72916 人の 1649 の映画に対する6段階評価のデータである．本実験ではそのうち 1800 の 1649 の映画に対してシミュレーションを行った．評価されている項目は全部で 107803 個あり欠損率は 96.4%である．ここではそのうち 87803 個をランダムに取り出し学習データとみなし，残りの 20000 個のデータで推定誤差を評価した．

両データともに同じ条件でシミュレーションを 5 回行い，その平均について評価した．

4．2 評価基準

提案手法を MAE と F 値で評価した．

MAE とは平均絶対誤差のことであり以下の式で表される．

$$\mathrm{MAE}=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \quad (式14)$$

ただし N はテストデータ数、p_i は予測値、q_i は正解値とする．

F 値とは推薦されるべきアイテムが正しく推定される割合を表しており以下の式で表される．ここでは評価値が 4 を超えるアイテムを推薦されるべきアイテムとした．

$$F=\frac{2PR}{P+R} \quad (式15)$$

ただし P（再現率）は

$$\frac{\text{推薦された適合アイテム数}}{\text{全テストデータ中の全適合アイテム数}} \quad (式16)$$

R（正解率）は

$$\frac{\text{推薦された適合アイテム数}}{\text{推薦されたアイテム数}} \quad (式17)$$

である．再現率は適合アイテムを漏れなく推薦できる度合い，正解率は適合アイテムだけを推薦できる度合いを表している．なお再現率と正解率はトレードオフ関係になっており，このふたつを同時に評価する評価基準として F 値は代表的な評価基準である．

4．3　評価結果と考察

4．3．1　相関係数法の評価と考察

表1，2に評価結果を示す．表1，2よりMovieLensデータではMAE，F値が従来手法に比べそれぞれ0.021，0.012向上した．EachMovieデータでもMAE，F値が従来手法に比べそれぞれ0.032，0.016向上した．これは嗜好に影響を与えているアイテムを考慮した「嗜好についての類似度」が改善されたためだと考えられる．実際にテストデータを予測する際に求めた類似度の式全てについて

$$\left|M_i + (M_{jz} - M_{j'}) - M_{iz}\right| \le 0.5 \cap C'_{ijz} - C_{ij} \ge 0 \quad (式18)$$

または

$$\left|M_i + (M_{jz} - M_{j'}) - M_{iz}\right| > 0.5 \cap C'_{ijz} - C_{ij} < 0 \quad (式19)$$

になる類似度の総数を調べた（ただしここでのM_{iz}は既知の値）．ここで（式18）は正解に寄与しているユーザの「嗜好についての類似度」を従来手法より増加している場合である．逆に（式19）は不正解に寄与しているユーザの「嗜好についての類似度」を従来手法より減少させている場合である．（式18）もしくは（式19）を満たしている類似度の式が全体の59%を占めており提案手法①が全体的に類似度の式を向上させていると考えられる．

表1　MovieLensデータの相関係数法での結果

	MAE	再現率	正解率	F値
従来手法	0.712	0.724	0.705	0.714
提案手法①	0.691	0.737	0.715	0.726

表2 EachMovieデータの相関係数法での結果

	MAE	再現率	正解率	F値
従来手法	0.859	0.654	0.750	0.699
提案手法①	0.827	0.671	0.764	0.715

4．3．2　二項関係学習法の結果と考察

表3，4に評価結果を示す．表3，4より従来手法の場合，MovieLensデータでは$\gamma=0.8$の時にMAEの値が，$\gamma=0.9$の時にF値の値が最も良くなった．EachMovieデータでは$\gamma=0.6$の時にMAEの値が，$\gamma=0.7$の時にF値の値が最も良くなった．提案手法②では従来手法が最も良いγを用いたものよりもMovieLensデータではMAE，F値が従来手法に比べそれぞれ0.016，0.009向上した．EachMovieデータでもMAE，F値が従来手法に比べそれぞれ0.021，0.017向上した．これは従来手法ではどのアイテムに対しても一定のウエイトを与えていたために，嗜好に影響を与えないアイテムが原因で嗜好の似ているユーザの類似度を下げていたのが改善されたためだと考えられる．

表3 MovieLensデータの二項関係学習法での結果

	MAE	再現率	正解率	F値
従来手法$\gamma=0.7$	0.766	0.756	0.698	0.726
従来手法$\gamma=0.8$	0.762	0.767	0.694	0.729
従来手法$\gamma=0.9$	0.763	0.782	0.688	0.732
提案手法②	0.744	0.775	0.710	0.741

表4：EachMovieデータの二項関係学習法での結果

	MAE	再現率	正解率	F値
従来手法$\gamma=0.5$	0.898	0.744	0.700	0.721
従来手法$\gamma=0.6$	0.894	0.751	0.699	0.724
従来手法$\gamma=0.7$	0.894	0.754	0.696	0.724
提案手法②	0.873	0.772	0.712	0.741

5むすび

本論文では「回答パターンの類似度」を用いて嗜好に影響を与えているアイテムに対してウエイトづけを行う手法を提案した．またシミュレーションを行い，従来手法に比べ，精度が向上する事を示した．

今後の課題として，どのようなデータの傾向があれば，「回答パターンの類似度」という概念を用いるのが有効なのかを考えたい．また人の嗜好というあいまいなものに対してどのような傾向があるのかを考え手法に反映させたい．

謝辞

著者の一人大島は，研究を進めるにあたり御討論，ご助言を頂いた湘南工科大学小林学先生に感謝します．

<参考文献>

[1] P.Resnick N.Iacovou M.Suchak P.Berstom "GroupLens:an open architechture for collaborative filtering of netnews", 1994, in Proc of CSCW94, pp.175-186

[2] A.Nakamura N.Abe, "Collaborative filtering using weighted majority prediction algorithms", 1997, in Proc.of ICML98, pp.395-403

[3] B.Sarwar G.Karypis J.Konstan Riedl, J, "Item-based Collaborative Filtering Recommender Algorithms.", 2001, in proc of the 10th international world wide web conference

[4]福原知宏"協調フィルタリングに関する研究動向" http://ai-www.aist-nara.ac.jp/doc/people/tomohi-f/Docs/cf_review.html,1998

[5] http://www.cs.umn.edu/Research/GroupLens

[6]http://www.research.digital.com/SRC/eachmovie