The 29th Symposium on Information Theory
and its Applications (SITA2006)
Hakodate, Hokkaido, Japan, Nov. 28 – Dec. 1, 2006

相関に基づいた共クラスタリングによる協調フィルタリング
Collaborative filtering by co-clustering based on correlation

高島 秀佳* 石田 崇* 平澤 茂一*
Hideyoshi TAKASHIMA Takashi ISHIDA Shigeiti HIRASAWA

Abstract— A recommendation system based on collaborative filtering needs large database for high estimated accuracy. However, the complexity of estimation increases as the size of the database takes a large value and it becomes more difficult to respond to users at once. In this paper, we propose a collaborative filtering algorithm by using co-clustering based on correlation with low complexity, and show the superiority of the proposed method by numerical experiments.

Keywords— collaborative filtering, recommendation system, co-clustering

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 [1]．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．

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 [1] は過去の購買履歴のデータから購買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代表的アルゴリズムに相関係数法 [1] がある．この手法はあるアイテムが好きか嫌いかという多数のユーザの多段階評価値について，評価値の相関係数を用いて特定のユーザの未評価の評価値を予測する．

このような推薦システムはユーザやアイテムの数が増え，入力される評価値が増えるほど推定の精度は向上する．しかし，それに伴い推定にかかる計算時間が増えユーザへのリアルタイムでの応答が難しくなってしまう．

そこで本研究では，相関係数法に対して共クラスタリング [3] を適用する手法を提案し，推定精度を大きく劣化させることなく計算時間を低減する手法を提案する．さらに，協調フィルタリングの分野で用いられているベンチマークデータ [5] に対して数値実験を行い提案手法の有効性を示す．

2 従来手法

2.1 協調フィルタリング

協調フィルタリング [1] とはあるアイテムに対するユーザの評価値を，そのユーザの別のアイテムに対する評価値と，他ユーザの評価値に基づいて予測する手法である．協調フィルタリングにおいて，データベースは行がユーザを表し，列がアイテムを表すユーザ-アイテム行列の形式で保持されている．この行列を $M=(M_{ix})$ とする．行列 M の (i,x) 要素 M_{ix} は i 番目のユーザ u_i の x 番目のアイテム v_x への評価値とする．ここで，M_{ix} はユーザが評価済みの場合には値をもち，ユーザが未評価の場合には欠損値となり値をもたない．通常，行列 M は欠損値を多く含み，ほとんどの評価値 M_{ix} には値が与えられない場合が多い．図 1 に行列 M の例を示す．

アイテム

ユーザ	v_1	v_2	v_3	v_4	…	v_x	…	v_k
u_1	4	2	5		…	4	…	1
u_2	1		2	4	…	5	…	4
u_3	4			2	…	3	…	2
⋮	⋮	⋮	⋮	⋮		⋮		⋮
u_i	1		2	5	…	M_{ix}	…	5
⋮	⋮	⋮	⋮	⋮		⋮		⋮
u_n	2		2	5	…	4	…	4

図 1: ユーザ-アイテム行列 M の例

協調フィルタリングの目的はユーザが未評価の要素に対してその評価値を予測することであり様々な手法が提案されている [1][4]．ここではその中でも多くの推薦システムに用いられている相関係数法 [1] を用いる．

2.2 相関係数法

相関係数法は協調フィルタリングの代表的な推定手法で，欠損値 M_{ix} に対して，推定対象ユーザ u_i の，推定対象アイテム v_x 以外のアイテムに対する評価値と，他ユーザの評価値に基づいて予測値 $\hat{M}_{ix}$ を算出する手法である．

[相関係数法における $\hat{M}_{ix}$ の推定法]

評価値の予測に用いる以下の値を定義する．M_i はユーザ u_i の評価値の平均値で，以下の式で定義する．

$$M_i = \frac{\sum_k M_{ik}}{\sum_k 1}. \quad (1)$$

ただし，ここでの $\sum_k$ は評価済のアイテムについてのみ計算する．また，評価対象ユーザ u_i とデータベース中のユーザ u_j の相関係数 w_{ij} を以下の式で定義する．

$$w_{ij} = \frac{\sum_k (M_{ik}-M_i)(M_{jk}-M_j)}{\sqrt{\sum_k (M_{ik}-M_i)^2 \sum_k (M_{jk}-M_j)^2}}, w_{ij} \in [-1,1] \quad (2)$$

ただし，ここでの $\sum_x$ は評価対象ユーザ i とデータ

* 〒 169-8555 東京都新宿区大久保 3-4-1 早稲田大学理工学部経営システム工学科 School of Science and Engineering,Waseda University Okubo 3-4-1, Sinjuku-ku, Tokyo, 169-8555 Japan.
E-mail: takasima@hirasa.mgmt.waseda.ac.jp

ベース中のユーザ j の両方で評価値を持つアイテムについてのみ計算する．以上の値を用いて未評価の評価値 M_{ix} を以下のように推定する．

$$\hat{M}_{ix} = M_i + \sum_j \frac{w_{ij}(M_{jx} - M_j)}{|w_{ij}|}. \quad (3)$$

ただし，ここでの $\sum_j$ はアイテム x を評価しているユーザのみについて計算する．

2.3 相関に基づく階層クラスタリング

クラスタリングを用いて相関係数法の計算時間を低減する手法が提案されている [2]．相関係数法ではユーザ間の類似度は相関係数で測られるので，クラスタリングもユーザ間の相関に基づいて行うが，行列 M は欠損を含むので k 平均法などはそのままでは適用できない．またクラスタの数も決めがたい．そこで，この手法では相関に基づく反復法 [2] を用いて次々にクラスタを二つに分割していく階層クラスタリングを用いている．

[相関に基づく反復法アルゴリズム]

Step0 全てのユーザを 1 つのクラスタとみなす．

Step1 行列 M より式 (2) を用いて算出されたユーザ u_i と u_j の相関係数 w_{ij} を求め，クラスタごとにユーザ-ユーザ行列 U を生成する．ここで，相関係数は共通に評価しているアイテムが無い場合計算できないので，その場合には未入力のままにしておく．

Step2 Step1 で得られた行列 U から各ユーザ間の相関係数 w'_{ij} を以下の式より算出する．

$$w'_{ij} = \frac{\sum_k (U_{ik} - U_i)(U_{jk} - U_j)}{\sqrt{\sum_k (U_{ik} - U_i)^2 \sum_k (U_{jk} - U_j)^2}}, w'_{ij} \in [-1, 1] \quad (4)$$

得られた相関係数 w'_{ij} を入力し，行列 U の値を更新する．

Step3 ユーザ間の相関係数が全てある閾値以上となったクラスタはそれ以上分割せず終了．そうでなければ Step4 へ．

Step4 行列の各要素が 1 と -1 に収束したとき，Step5 へ．そうでなければ step2 に戻る．

Step5 各ユーザ間の相関が 1 となったユーザのグループが 2 つできるのでそれぞれをクラスタとみなし，2 つのクラスタに分割し Step1 へ．

最初に得られる行列 U は表 1 のように未入力の要素を含んでいるが，ユーザ間の相関係数を計算し値を更新することで，表 2 のように未入力の要素を含まない行列 U が得られる．さらに反復法を用いることにより表 3 のような行列 U が得られることとなる．このとき，クラスタの分割が起こる．分割されたクラスタに対して同様にアルゴリズムを繰り返す．

表 1: 行列 U の例

	u_1	u_2	u_3	u_4
u_1	1	0.8	-0.4	0.2
u_2	0.8	1		-0.5
u_3	-0.4		1	0.9
u_4	0.2	-0.5	0.9	1

表 2: 反復 1 回目の行列 U

	u_1	u_2	u_3	u_4
u_1	1.00	0.70	0.15	0.28
u_2	0.70	1.00	-0.61	-0.38
u_3	0.15	-0.61	1.00	0.93
u_4	0.28	-0.38	0.93	1.00

表 3: 収束状態

	u_1	u_2	u_3	u_4
u_1	1	1	-1	-1
u_2	1	1	-1	-1
u_3	-1	-1	1	1
u_4	-1	-1	1	1

2.4 従来手法の問題点

相関係数法を用いて評価値 M_{ix} に対する推定精度を高めるためには，データベース中のユーザやアイテム，入力される評価値を増やす必要がある．しかし，データベースが大きくなると推定にかかる計算時間が増加し，ユーザへのリアルタイムでの応答が難しくなってしまう．そこで [2] のようにユーザのクラスタリングを用いて計算時間を低減する方法が提案されている．

一般に協調フィルタリングのデータはユーザ - アイテム行列で表現されているが，ユーザ数よりもアイテム数の方が多く，クラスタリングを行う際にはアイテムに対してもクラスタリングを行うことで，より効果の高いクラスタリングが実現できると考えられ，行列に対して行-列双方をクラスタリングする手法として共クラスタリング [3] がある．

そこで，提案手法では推薦システムに広く使われている相関係数法に特化した相関に基づく共クラスタリング手法を提案する．クラスタ内に一つでも欠損していない評価値をもつ場合にはクラスタ内の全ての評価値が値を持つことになり，未評価の評価値にも値が与えられ，欠損値を減らすことができる．共クラスタリングを用いた場合，ユーザのみをクラスタリングする場合よりも 1 クラスタに含まれる評価値が増え，その結果さらに欠損値を減らすことができ，密な行列を得ることができる [3]．

3 提案手法

本節ではユーザ u_i のアイテム v_x に対する評価値 M_{ix} を予測する問題を考える．本節では相関係数法の計算時間低減のため，相関係数法に基づきユーザとアイテムを共クラスタリングする手法を提案する．以下にそのアルゴリズムを示す．

[提案アルゴリズム]

Step1 相関に基づく反復法を用いてユーザ-ユーザ行列 U を生成し，ユーザをクラスタリングすることでユーザクラスタ c_y を生成する．

Step2 相関に基づく反復法のユーザをアイテムに置き換えてアイテム-アイテム行列 I を生成し，アイテムをクラスタリングすることでアイテムクラスタ c_z を生成する．Step1 の結果を用いてユーザ-アイテムクラスタ c_{yz} を生成する．

Step3 得られた各クラスタ c_{yz} 内の評価値の平均を算出し，その平均値をクラスタの評価値 $M_{c_{yz}}$ とする．ただし，クラスタ内に評価済の評価値が 1 つもないときは空欄のままとする．

Step4 評価値 M_{ix} の推定を行うとき，評価値 M_{ix} を含むクラスタに評価値 $M_{c_{yz}}$ が与えられているとき，以下の式のように評価値 $M_{c_{yz}}$ を評価値 M_{ix} の推定値とする．

$$\hat{M}_{ix} = M_{c_{yz}}, c_{yz} \in M_{ix} \tag{5}$$

Step5 評価値 M_{ix} を含むクラスタ cyz に評価値 $M_{c_{yz}}$ が与えられていないとき，クラスタリングされたデータベースに対して以下のように相関係数法を用いて推定を行い，式 (7) より得られた推定値 $\hat{M}_{c_{yz}}$ を評価値 M_{ix} の推定値とする．

$$M_{c_y} = \frac{\sum_k M_{c_{yk}}}{\sum_k 1}. \tag{6}$$

ただし，$\sum_k$ はクラスタ c_{yk} が評価済のときのみ計算する．

$$w_{c_y c_a} = \frac{\sum_k (M_{c_{yk}} - M_{c_y})(M_{c_{ak}} - M_{c_a})}{\sqrt{\sum_k (M_{c_{yk}} - M_{c_y})^2 \sum_k (M_{c_{ak}} - M_{c_a})^2}}, \quad w_{c_y c_a} \in [-1, 1]. \tag{7}$$

ただし，$\sum_k$ はクラスタ c_{yk} とクラスタ c_{ak} がともに評価済のときのみ計算する．

$$\hat{M}_{ix} = M_{c_y} + \sum_k \frac{w_{c_y c_a}(M_{c_{kz}} - M_{c_k})}{|w_{c_y c_a}|}. \tag{8}$$

ただし，$\sum_k$ はアイテム x を含むクラスタ c_{kz} が評価値をもつときのみ計算する．

本手法では Step1 から Step3 は通常，事前にオフラインで行割れるものと仮定する．本手法を用いることで，データベース中のユーザ数及びアイテム数がクラスタ数に減らされたことになり，オンラインでの推定要求に対する処理が高速化される．

4 数値実験による評価

4.1 利用データ

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 Movie Lens データ [5] を用いた．Movie Lens データは 943 人の 1682 本の映画に対する 5 段階評価値データである．評価されている項目は全部で 100,000 個ありデータの欠損率は 93.7 % である．

本実験ではユーザ 100 人のアイテム 500 個に対する評価の部分をデータベースとして用いた．

4.2 実験方法

本手法を用いることにより推定にかかる計算時間は低減されるが推定精度は下がると考えられる．

そこで，本実験では，

(従来手法 1) クラスタリングを行わなかった場合の相関係数法

(従来手法 2) ユーザのみにクラスタリングを行った場合の相関係数法

(従来手法 3) ユーザ数とアイテム数がクラスタ数と同数の場合の相関係数法

の 3 つとの比較を行う [2].

1) 計算時間の比較

提案手法はユーザ数とアイテム数がクラスタ数と同じ行列 M に対して相関係数法を用いた場合と計算時間が同じであるので，提案手法と従来手法 3 の計算時間は同じであると考えられる [2]．そこで，計算時間の比較は，提案手法と従来手法 1，従来手法 2 との間でのみ行う．これにより共クラスタリングを用いたことによる本手法の計算時間低減の効果を示す．

2) 推定精度の比較

推定精度については提案手法と従来手法 1 ～ 3 の間で行う．これにより，行列 M をクラスタリングしたことにより推定精度は下がってしまうが，共クラスタリングを用いているほうが，ユーザ数とアイテム数がクラスタ数と同数の行列 M に対して相関係数法を用いた場合よりも推定精度は上であることを示す [2].

また，本実験では「All but 1」方式を採用した [?].「All but 1」方式は全評価データのうち 1 つをランダムに選びその評価値をマスクし，残りの評価データからその評価値を予測し評価するという方式である．

ここではある M_{ix} に対して推定を 1 回行うことを 1 回の実験とみなし，$N = 100$ 回 の実験の平均をとった．

4.3 推定精度の評価方法

提案手法の推定精度の評価には，平均絶対誤差 MAE と F 値 [4] を用いた．MAE は値が小さいほどよく，F 値は値が大きいほどよい．ただし，ここでは評価値が 4 以上のアイテムを推薦されるべき適合アイテムとし，適合アイテムが正しく推定される割合を表す評価基準として R（再現率）と P（正解率）がある．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おり，以下の式で計算される．

$$R = \frac{\text{推薦された適合アイテム数}}{\text{データ中の全適合アイテム数}}, \tag{9}$$

$$P = \frac{\text{推薦された適合アイテム数}}{\text{推薦されたアイテム数}}. \tag{10}$$

ここで再現率と正解率はトレードオフの関係になって

おり，今回は両方を同時に評価する評価基準として F 値を用いる．F 値は以下の式で計算される．

$$F = \frac{2PR}{P+R}. \tag{11}$$

4.4 実験の結果と考察

4.4.1 計算時間の比較

相関に基づく反復法アルゴリズムの閾値を 0.9 として，ユーザ-アイテム行列 M に対して共クラスタリングを行った結果，ユーザは 10 個，アイテムは 62 個のクラスタに分割された．クラスタリングに要した時間は 1.3GHz celeron M プロセッサ，1GByte メモリで 24 秒であった．

クラスタリングはオフラインで行うものとしているので，比較対象となるのは 1 回の推定ごとにかかる計算時間である．図 2 に実際に計測した計算時間を示す．

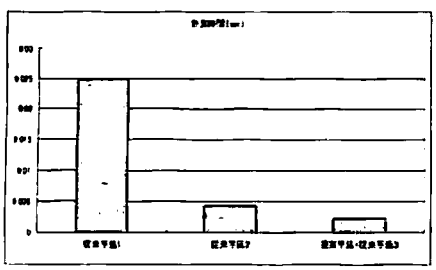

図 2: 計算時間の比較

従来手法 1 と比較すると提案手法はクラスタリングを行うことで，行列 M のユーザ数とアイテム数がクラスタ数と同数である場合と計算時間は同じである．つまり，行列 M を小さくしたのと同じ効果が得られるので計算時間も短縮されている．

また，従来手法 2 と比較してみると，アイテムに対してもクラスタリングを行っている分，計算時間も半分程度に低減されていることが分かる．

4.4.2 推定精度の比較

提案手法と従来手法の MAE を比較した結果を図 3 に，F 値を比較した結果を図 4 に示す．

従来手法 1 と比較すると，行列 M に対してクラスタリングを行った弊害として MAE も F 値も下がっていることが分かる．次に従来手法 2 と比較すると，MAE は若干下がっているものの，F 値ではほとんど差がないことが分かる．F 値に差がないということは，推薦システムに提案手法を適用した場合，従来手法 2 とほぼ同じ精度で適合アイテムを推薦できるということである．推薦システムにおいて必要な要件は「推薦すべき適合アイテムをきちんと推薦出来ること」であると考えられるので提案手法は推薦システムにおいて従来手法 2 と同じ性能を発揮できると考えられる．

最後に従来手法 3 と比較すると，提案手法はユーザ 10 人，アイテム 62 個の行列 M に相関係数法を用いた場合と比べると，計算時間は同じで推定精度は高いという結果が得られた．

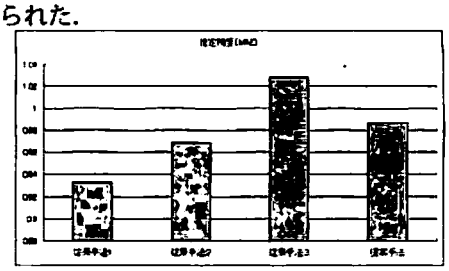

図 3: MAE の比較

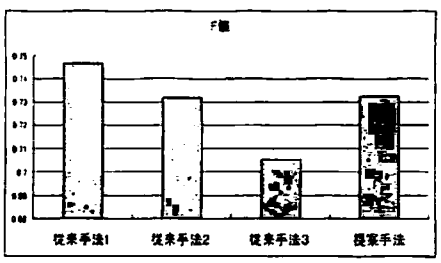

図 4: F 値の比較

5 おわりに

本研究では，相関に基づく共クラスタリングを用いて計算時間を低減させる協調フィルタリング手法を提案した．また数値実験を行い従来手法と比較し，本手法の有効性を示した．

今後の課題として以下のようなものがあげられる．

(1) オンラインでのクラスタの更新の方法

本手法ではオフラインでクラスタの更新を行っており，更新のためには一時的にシステムを停止する必要がある．今後はシステムを停止させずにオンラインでクラスタを更新する手法を考える必要がある．

(2) 推定精度の向上

本研究では計算時間の低減には成功したが，推定精度は下がってしまった．今後は推定精度をできるだけ下げないような方法を考える必要がある．

参考文献

[1] John S. Breese, David Heckerman, carl Kadie, "Empirical Analysis of Predictive Algorithms for collaborative Filtering," *in Proc. of the 14th conference on Uncertainty in Artificial Intelligence*, pp.43-52, 1998.

[2] 井上 光平，浦浜 喜一，"人のクラスタリングによる協調フィルタリングの高速化"，電子情報通信学会論文誌，J87-D-II, No.12, pp.2700-2702, 2001.

[3] Thomas George, Srujana Merugu, "A Scalable collaborative Filtering Framework based on co-clustering," *in Proc. of the 5th IEEE International conference on Data Mining* ,2005.

[4] 大島敬志，"観測値の類似度を考慮した協調フィルタリング，" 早稲田大学大学院理工学研究科修士論文, 2003.

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