1F4-05

なんとなく協調フィルタリング — 複数の順序応答に基づく推薦

Nantonac Collabortive Filtering — Recommendation Based on Multiple Order Responces

神嶌 敏弘[*1]
Toshihiro Kamishima

[*1]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

A recommender system suggests the objects expected to be preferred by the users. Recommender systems use collaborative filtering to recommend objects by summarizing the preferences of people who have the similar preference patterns to the target user. Traditionally, these preference patterns are represented by making use of rating scores. We had developed recommendation methods using order responses instead of rating scores, and had shown advantages of using orders. However, this framework bear a problem that the number of objects per user is limited. To overcome this limitation, we extend this framework so that multiple orders can be collected from each user. We propose several methods to deal with these multiple order responses.

1. はじめに

推薦システム（Recommender System）[Ben Schafer 01] は，利用者が好むであろうアイテム見つけ出すもので，協調フィルタリング（Collaborative Filtering; CF）は,「口コミ」の原理を用いる推薦システムの一手法である．この CF の実現には利用者の嗜好の度合いを計測する必要が生じるが，ほとんどの CF では Semantic Differential 法 (SD 法) [Osgood 57] が用いられる．この方法では，利用者の嗜好を「好き」と「嫌い」という対義語で両端を表し，その間を 1〜5 などの数字で等間隔に区切った物差しを使って計測する．こうして得られたスコアを間隔尺度として扱うために「ある一つの物差しの目盛は間隔は等しい」と,「全ての物差しの全長は等しい」という二つの仮定が必要になる．しかし，これらの仮定を満たすようなスコアを得るのは一般には難しい．また，心理学的要因によるスコアのずれを生じるという問題もある．そこで，我々は，利用者の嗜好を順序によって提示する順位法を用いた CF である「なんとなく協調フィルタリング」[Kamishima 03a, 神嶌 04a] なる枠組みを提案し，SD 法よりも順位法による嗜好パターンの計測によって，推薦の精度を改善できることを示した．

しかし，順位法には，利用者は一度にせいぜい 10 個程度の対象しか整列できないという制限がある．従来のなんとなく CF では利用者一人に一つの順序応答しか許していないため，集積できる嗜好データの量が大きく制限される問題があった．そこで，利用者一人につき複数の順序応答を許すようになんとなく CF の枠組みを拡張する．本研究ではその第一段階として，システムを利用中の活動利用者の順序応答は一つだが，データベース中の標本利用者の応答順序は複数でもよい場合について，従来の手法を拡張し，その性能を実験により検証する．

以下，2. 節ではなんとなく CF の枠組みを，3. 節ではその手法を，4. 節では実験結果を示し，最後に 5. 節にまとめを述べる．

2. なんとなく協調フィルタリング

従来のなんとなく CF の枠組みと，今回の拡張した枠組みについて形式的に述べる．

連絡先: 神嶌敏弘，http://www.kamishima.net/

2.1 単単なんとなく CF

協調フィルタリングは，ある特定の利用者（活動利用者）の嗜好を，他の利用者（標本利用者）の嗜好情報を集積したデータベース（利用者 DB）から予測する問題である．なんとなく CF とは形式的には以下のような問題である．推薦や評価される対象を x_j，その全集合を $X^*=\{x_1,\ldots,x_{|X^*|}\}$ とする．システムが対象集合 X_i を利用者 i に提示すると，利用者はこれらの対象を嗜好の度合いによって整列した順序 $O_i=x_1\succ x_2\succ\cdots\succ x_{|X_i|}$ を返す．順序 O_i を構成する対象集合は X_i の他に $X(O_i)$ とも表記する．また順序中の対象 x_j の添え字 j は，順序中 j 番目の対象ではなく X^* 中の一つの対象を一意に指定するインデックスであることに注意されたい．順位 $r(O_i,x_j)$ は，対象 x_j の順序 O_i 中の位置を示す基数である．例えば，順序 $O_i=x_1\succ x_3\succ x_2$ について $r(O_i,x_2)$ は 3 である．標本利用者が整列した順序の集合 $D_S=\{O_1,\ldots,O_{|D_S|}\}$ を利用者 DB とし，活動利用者の順序応答を O_a とする．このとき，O_a と D_S から，任意の対象集合に対する活動利用者の嗜好順序を予測し，その上位の対象を推薦する問題がなんとなく CF である．この従来のなんとなく CF を活動利用者も標本利用者も一人につき一つの順序応答を返すので，ここでは単単なんとなく CF と呼ぶ．

さらに順序に関するいくつかの概念を述べておく．全ての対象を含む順序，すなわち $X_i=X^*$ なる O_i は完全順序（complete order），そうでないとき，不完全順序（incomplete order）と呼ぶ．二つの順序 O_1 と O_2 について，$x_a,x_b\in X_1\cap X_2, x_a\neq x_b$ なる対象 x_a と x_b があるとき，x_a と x_b について O_1 と O_2 が同順（concordant）であるとはこれら二つの対象が同じ順に順序中現れることで，形式的には次の条件が満たされることである．

$$(r(O_1,x_a)-r(O_1,x_b))(r(O_2,x_a)-r(O_2,x_b))\geq 0$$

また，そうでないとき非同順（discordant）であるという．O_1 と O_2 について，$x_a,x_b\in X_1\cap X_2, x_a\neq x_b$ なる全ての対象の対について同順なら，O_1 と O_2 は同順であるいう．距離 $d(O_a,O_b)$ は同じ対象で構成される二つの順序，すなわち，$X_a=X_b(\equiv X)$ を満たす順序間で定義される．Spearman の距離 $d_S(O_a,O_b)$ [Marden 95] はよく用いられる距離の一つで，順位の差の二乗和で定義される．

$$d_S(O_a,O_b)=\sum_{x_j\in X}\left(r(O_a,x_j)-r(O_b,x_j)\right)^2 \qquad (1)$$

この距離を $[-1,1]$ の範囲に正規化したものが Spearman の順位相関 ρ で次式で，定義される．

$$\rho = 1 - 6 \times d_S/(|X|^3 - |X|). \quad (2)$$

2.2 単復なんとなく CF

1. 節で述べたように，認知能力の限界のため，順位法では利用者は一度にせいぜい 10 個程度の対象しか整列することができない．そのため，利用者一人について一つの順序応答しか許さない単単なんとなく CF では，利用できる嗜好データの量がどうしても制限されてしまう問題がある．そこで，活動利用者は従来どおり一つの順序応答しか返せないが，標本利用者は一人につき複数の順序応答を返すことを許すように拡張したものを単複なんとなく CF と呼ぶ．標本利用者 i は対象集合 $X_{i1} \ldots X_{i,|Q_i|}$ をシステムより提示され，それぞれの集合に対して順序応答 $O_{i1} \ldots O_{i,|Q_i|}$ を返す．この順序応答の集合を Q_i で表す．利用者 DB はこの順序応答集合の集合 $D_S = \{Q_1, \ldots, Q_N\}$ で表され，この D_S と活動利用者の順序応答 O_a から，活動利用者の嗜好を予測するのが，単複なんとなく CF である．

3. 手法

従来の単単なんとなく CF 手法を修正し単複なんとなく CF で利用できるようにする．以前の研究 [Kamishima 03a, 神嶌 04a] では，GroupLens[Resnick 94] の方法に基づく単純相関法とクラスタリングを利用する方法を提案した．ここでは，それぞれについて二種類ずつ，単複なんとなく CF への拡張手法を示す．

3.1 単純相関法

まず，単単なんとなく CF での単純相関法 (SCR) を示す．これは，GroupLens の方法を順序に適用できるように修正したもので，順位をそのまま，GroupLens の手法におけるスコアと置き換えただけである．CF では，活動利用者と標本利用者の嗜好の類似性を測る必要があるが，GoupLens の方法ではこれを通常の Pearson 相関で測る．だが，なんとなく CF では嗜好は順序で表されているので，かわりに，式 (2) の順位相関 ρ を用いる．ここで，順位相関は同じ対象で構成される順序の間でしか定義されない．しかし，活動利用者と標本利用者の応答順序を構成する対象集合は等しくない場合がほとんどである，すなわち，$X(O_a) \neq X(O_i)$．そこで，ここでは文献 [Kamishima 04b, 神嶌 05] の欠損した対象の順序を補完する手法を用い順位相関を計算する．ただし，補完に用いた中心順序は 3.3 節の方法で計算した．補完後の順序の Spearman の ρ を R_{ai} で表す．活動利用者の対象 x_j についての嗜好は次式で予測する．

$$\hat{r}_{aj} = \frac{\sum_{i \in \tilde{U}_j} R_{ai} \; r(O_i, x_j) - \bar{r}_i}{\sum_{i \in \tilde{U}_j} |R_{ai}|}, \quad (3)$$

ただし，$X_{ai} = X_a \cap X_i$，$\bar{r}_i = (\sum_{x_j \in X_{ai}} r(O_i, x_j))/|X_{ai}|$，$\tilde{U}_j$ は対象 x_j を応答順位に含む利用者を表す番号の集合，すなわち，$\{i | O_i \in D_S, x_j \in X_i\}$．対象をこの式の値で昇順に整列し，上位に順位付けされたものを推薦する．

この単純相関法を，単複なんとなく CF に適用できるように修正した方法を二種類示す．一つは単純に，Q_i が同じ利用者から得られた応答順序であるという情報を無視し，$Q_1, \ldots, Q_N$ なる集合を併合した集まり $D_S^{\text{merge}} = \{O_{11}, \ldots, O_{1,|Q_1|}, O_{21}, \ldots, O_{N,|Q_N|}\}$ を生成する．この D_S^{merge} を利用者 DB と考えて，単単なんとなく CF の単純相関法を適用したものを，併合単純相関法（MGS）と呼ぶ．もう一つは，応答順序集合 $Q_1, \ldots, Q_N$ それぞれについて，その中心順序を 3.3 節の方法で求める．それを集めた $D_S^{\text{center}} = \{\bar{O}_1, \ldots, \bar{O}_N\}$ なる集合を利用者 DB と考えて，単単なんとなく CF の単純相関法を適用するものを，中心順序単純相関法（COS）と呼ぶ．

3.2 クラスタリング法

クラスタリング法（CLS）は事前に利用者 DB をクラスタリングしておき，活動利用者の応答順序が所属するクラスタの情報を利用して，嗜好を予測する方法である．クラスタリングには文献 [Kamishima 03b, 神嶌 03c] の k-o'means-TMSE 法を修正した k-o'means-MMER 法を用いた．これは，中心順序を Thustone の方法ではなく，3.3 節の補正平均期待順位法で求める点のみが異なる．単単なんとなく CF で推薦を行うには，推薦の前に，利用者 DB D_S を k 個のクラスタ $\{C_1, \ldots, C_k\}$ に分割する．活動利用者の順序 O_a が与えられると，各クラスタの中心順序と O_a の距離 $1 - \rho(\bar{O}_{C_j}, O_a)$ を求めて，活動利用者と最も類似した嗜好をもつクラスタ C^* を見つける．そのクラスタの中心順序 $\bar{O}_{C^*}$ で上位にある対象を推薦する．

単純相関法では，O_a を受け取ったあと，$|D_S|$ 個の標本利用者の応答順序との比較が必要になるので，計算量は $O(|D_S|)$ と多い．一方，クラスタリング法では，事前にクラスタリングを済ませておけば k 個の中心順序と比較すればよいので計算量は $O(k)$ と抑えられる．しかし，我々の今までの実験では，単純相関法の方がより精度の高い推薦が実現できている．

このクラスタリング法を，単複なんとなく CF に適用できるように修正した方法を二種類示す．一つは MGS 法と同様に単純に応答順序を併合した D_S^{merge} を生成し，単単なんとなく CF のクラスタリング法を適用する併合クラスタリング法（MGC）である．もう一つはクラスタリング手法に，Wagstaff の MUST リンク制約 [Wagstaff 01] を導入する手法である．MUST リンク制約とは MUST リンクで結合された対象は必ず同じクラスタに分類されなければならないという制約である．ここでは同じ Q_i に由来する応答順序 $O_{i1}, \ldots, O_{i,|Q_i|}$ の間全てに MUST リンクがあると考える．この制約が満たされるように，k-o'means を実行中，応答順序をクラスタへ割り当てるステップで，クラスタの中心順序と Q_i 中の各順序との距離の総和が最小になるクラスタへ Q_i 中の全順序を割り当てる．こうして得られたクラスタを用いて推薦をする手法を MUST リンク・クラスタリング法（MLC）と呼ぶ．

3.3 補正平均期待順位法

ここでは中心順序を求める方法について述べる．順序集合 $S = O_1 \ldots O_N$ が与えられたとき，中心順序とは対象集合 $X(\bar{O}_S) = \cup_S X(O_i)$ で構成される順序で，次の条件を満たすものである [Marden 95].

$$\bar{O} = \arg \max_{O: X(O) = \cup_S X(O_i)} \sum_{O' \in S} d(O', O) \quad (4)$$

ただし，$d(O', O)$ は式 (1) の d_S のような順序間の距離で，O のうち O' に含まれない対象を無視して測ったものとする．しかし，この中心順序を求める問題は，一般には線形順序付け (linear ordering) 問題 [Grötschel 84] となり，NP 困難になる．そのため，我々の従来の研究では，Thurstone の比較判断の法則 [Thurstone 27, Marden 95] なる確率モデルで最小二乗エラー基準で推定する方法 [Mosteller 51] で近似解を求め

た．しかし，中心順序の長さに対して 2 乗の計算量が必要であり，より高速な方法が必要になった．そこで，本論文では以下の補正平均期待順位法を考案した．まず，X^* で構成される完全順序 O^* から，対象が均一な確率で欠損し不完全順序 O_i が観測されたとき，O_i 中の対象 x_j の O^* 中での期待順位は次式となることが知られている [Arnold 92].

$$\mathrm{E}[r(O^*, x_j)|O_i] = r(O_i, x_j) \times \frac{|O^*|+1}{|O_i|+1}$$

集合 S 中で対象 x_j を含む順序の集合を $S(x_j)$ で表すとき，補正平均期待順位を次式で定義する

$$\mathrm{MMER}(x_j) = \frac{0.5(|O^*|+1) + \sum_{O_i \in S(x_j)} \mathrm{E}[r(O^*, x_j)|O_i]}{|S(x_j)|+1} \quad (5)$$

$0.5(|O^*|+1)$ の項は事前の期待順位として，全ての対象に順序の中心の順位を $1/|O^*|!$ 個ずつ割り当てたとみなした補正項である．この $\mathrm{MMER}(x_j)$ を全ての $x_j \in X(\bar{O})$ について求め，これらの対象を $\mathrm{MMER}(x_j)$ について昇順に整列した順序を中心順序とするのが補正平均期待順位法である．この平均補正順位は S 中の全順序を一度だけ走査すればよいので，$O(\sum_S |X(O_i)|)$ 程度の時間で計算でき，中心順序はそれに整列時間 $O(|X(\bar{O})| \log |X(\bar{O})|)$ を加えた時間で計算できる．なお，同順位が生じた場合には同順位になる対象に，それらの対象に割り当てられる順位の平均値である midrank 値 [Marden 95] を一律に割り当てた．

4. 実験

3. 節の手法を文献 [神嶌 04a] でなんとなく CF の検証に用いた寿司の嗜好データに適用する．このデータでは，利用者はそれぞれ大きさが 10 の対象集合 X^A と X^B についてその嗜好順序を示している．このうち，X^A に対する嗜好順序 O^* を隠し，それを X^B への嗜好順序データを基に予測した $\hat{O}$ を求め，$\hat{O}$ と O^* の間の順位相関 ρ によって評価を行う．X^B に対する D 個の順序を活動利用者用と標本利用者用に分け，10 分割交叉確認を実施した．単単なんとなく CF の実験では標本利用者用の順序をそのまま利用したが，単複なんとなく CF では複数の標本順序が必要になる．そこで長さ 10 の順序からランダムに 7 個か 5 個の対象をサンプリングし，元の標本順序と同順になるように整列することで Q_i を生成した．なお，クラスタリング法では 20 回の試行でクラスタリングの目的関数を最良にする結果を選び，さらにクラスタ数を 2〜10 まで変えて，予測 ρ が最良の結果を示した．

最初に，同じ利用者からの応答順序の長さの違いによる変化を調べるため，長さ 7 と 5 の順序を，それらを構成する対象に重複を許して 2 個ずつ生成した結果を図 1 に示す．縦軸は上記の ρ で大きな値ほどよい推定であり，横軸はデータ中の利用者 DB 中の利用者数で，利用者数が多いほど嗜好の推定が容易になる．図の (a) は単純相関法を (b) はクラスタリング法を示している．単単なんとなく CF で長さ 10 と 5 の結果を示したのが SCR10 や SCR5 で，それぞれ予測精度の上限と下限の基準として参考にされたい．単純相関法では MGS と COS いずれでも，長さ 5 の単複なんとなく CF で，ほぼ推定精度の上限である SCR10 と同等の精度が得られており，複数の順序応答を利用する利点が示されている．また，MGS と COS の比較では大きな性能差は見られない．COS は事前に中心順序を計算しておくことで，MGS より高速に推薦を実行できるため，COS の方が優れているといえる．クラスタリング法でも，複数の順序応答を用いることで CLS5 より大きく推薦精度が改善されており，複数応答の導入は有効である．MGC と MLC を比較すると，応答順序が同じ利用者に由来するという情報を明示的に利用できるため，後者の方が推定精度が優れていた．

次に，図 2 に，長さ 5 の応答順序を 2 個用いたとき，それら 2 個の順序に含まれる対象に重複がある場合とそうでない場合の結果を示した．重複があると二つの順序の統合の精度が向上する．一方，重複がない方が多くの対象についての嗜好情報を同じ数の順序から得られる．グラフ中 ov は重複がある場合，dj は重複がなく互いに素な場合を表す．単純相関法では，差はわずかだが，どちらの手法も重複を許さず，できるだけ多くの対象について嗜好情報を獲得する戦略の方が有利なようである．これは，COS 法で複数の応答順序の中心順序を求めるとき，重複があるとしてもわずかなので，中心の推定精度に大きくは影響しないためと考えられる．クラスタリング法では，MGC では重複のある方が，MLC では重複のない方が若干有利である．MGC は，どの応答順序が同じ利用者由来であるという情報を破棄しているが，重複した対象は，同じ利用者の応答順序では同順になり同じクラスタに分類されやすくなることが働いて，この破棄した情報がある程度は補完され精度が改善したと思われる．一方，MLC 法ではすでにそのような情報は考慮されているので，より多くの対象についての嗜好情報がある，重複のない場合が有利になったと考えられる．

5. まとめ

従来のなんとなく CF の枠組みでは，利用者一人につき一つの順序応答しか想定していなかった．そのため，各利用者から収集できる嗜好の情報量が制限される問題があった．そこで，本研究では，標本利用者について一人につき複数の応答順序がある場合に拡張した単複なんとなく CF について論じた．単複なんとなく CF のためのいくつかの修正手法を提案し，その有効性を確かめた．今後は活動利用者も複数の応答順序を許す複複なんとなく CF について取り組みたい．

謝辞： 本研究は科研費 14658106 と 16700157 の助成を受けた．

参考文献

[Arnold 92] Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N.: *A First Course in Order Statistics*, John Wiley & Sons, Inc. (1992)

[Ben Schafer 01] Ben Schafer, J., Konstan, J. A., and Riedl, J.: E-Commerce Recommendation Applications, *Data Mining and Knowledge Discovery*, Vol. 5, pp. 115–153 (2001)

[Grötschel 84] Grötschel, M., Jünger, M., and Reinelt, G.: A Cutting Plane Algorithm for the Linear Ordering Problem, *Operations Research*, Vol. 32, No. 6, pp. 1195–1220 (1984)

[Kamishima 03a] Kamishima, T.: Nantonac Collaborative Filtering: Recommendation Based on Order Responses, in *Proc. of The 9th Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 583–588 (2003)

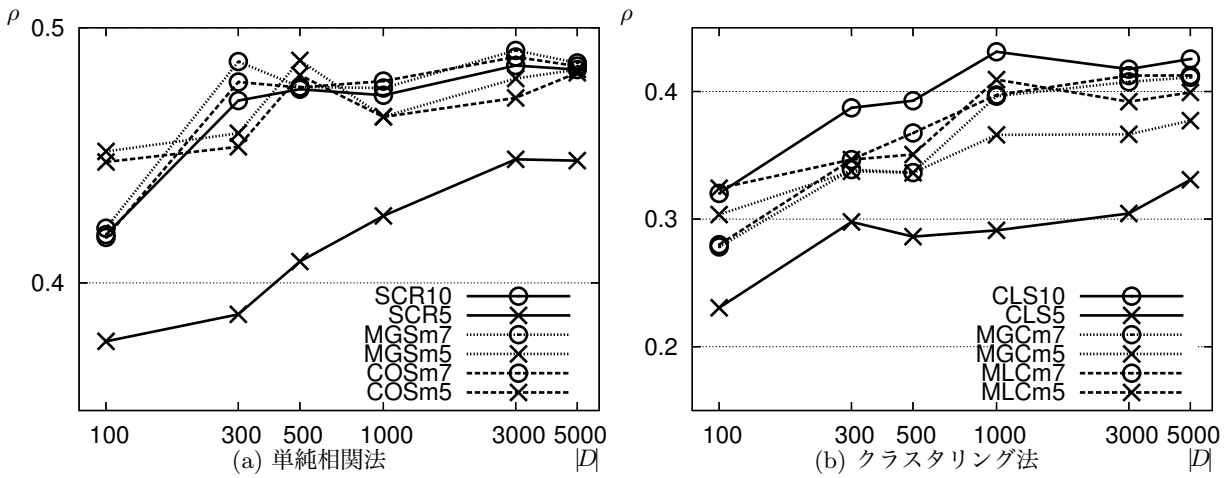

図 1: 標本順序応答を 2 個ずつ用いて，長さが 7 と 5 場合の比較

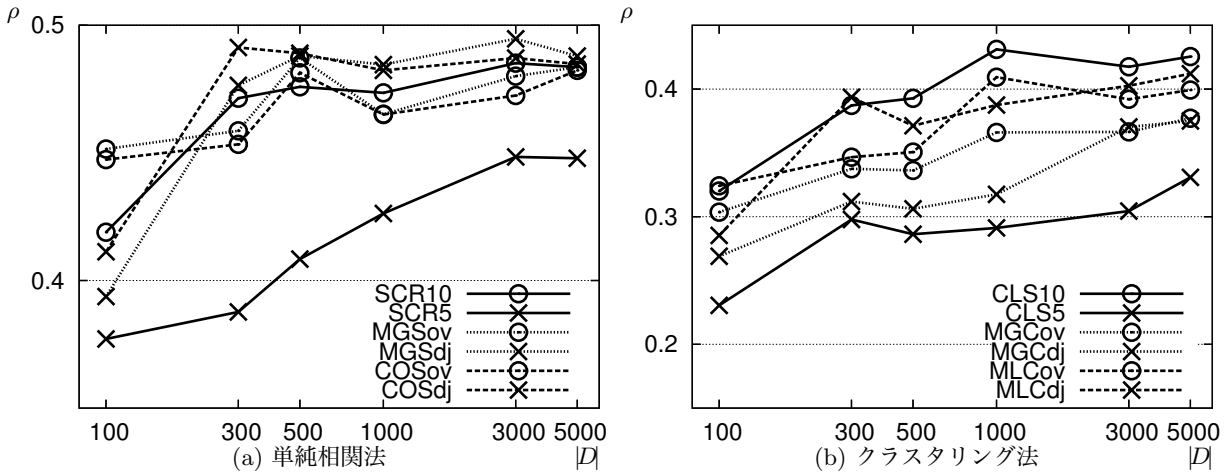

図 2: 長さが 5 の標本順序応答を 2 個用いて，それらに重複がある場合とない場合の比較

[Kamishima 03b] Kamishima, T. and Fujiki, J.: Clustering Orders, in *Proc. of The 6th Int'l Conf. on Discovery Science*, pp. 194–207 (2003), [LNAI 2843]

[神嶌 03c] 神嶌 敏弘, 藤木 淳：順序のクラスタリング — 順序平均の最適性について, 電子情報通信学会技術研究報告, PRMU 2003–83 (2003)

[神嶌 04a] 神嶌 敏弘：なんとなく協調フィルタリング — 順序応答に基づく推薦, 人工知能学会研究会資料, SIG–KBS–A304-37 (2004)

[Kamishima 04b] Kamishima, T. and Akaho, S.: Filling-in Missing Objects in Orders, in *Proc. of The 4th IEEE Int'l Conf. on Data Mining*, pp. 423–426 (2004)

[神嶌 05] 神嶌 敏弘, 赤穂 昭太郎：順序中の欠損対象の補完, 人工知能学会研究会資料, SIG–KBS–A405-13 (2005)

[Marden 95] Marden, J. I.: *Analyzing and Modeling Rank Data*, Vol. 64 of *Monographs on Statistics and Applied Probability*, Chapman & Hall (1995)

[Mosteller 51] Mosteller, F.: Remarks on the Method of Paired Comparisons: I — The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations, *Psychometrika*, Vol. 16, No. 1, pp. 3–9 (1951)

[Osgood 57] Osgood, C. E., Suci, G. J., and Tannenbaum, P. H.: *The Measurement of Meaning*, University of Illinois Press (1957)

[Resnick 94] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, in *Proc. of The Conf. on Computer Supported Cooperative Work*, pp. 175–186 (1994)

[Thurstone 27] Thurstone, L. L.: A Law of Comparative Judgment, *Psychological Review*, Vol. 34, pp. 273–286 (1927)

[Wagstaff 01] Wagstaff, K., Cardie, C., Rogers, S., and Schroedl, S.: Constrained K-means Clustering with Background Knowledge, in *Proc. of The 18th Int'l Conf. on Machine Learning*, pp. 577–584 (2001)