情 報 学 基 礎 61−25
自 然 言 語 処 理 142−25
（2001. 3. 6）

協調的フィルタリングにおける デフォルト投票値の効果的な設定方式

村松茂樹 井ノ上直己 帆足啓一郎 橋本和夫

(株)KDD 研究所

相関係数を重みとし，他の利用者の評価の重み付け和に基づいた利用者の評価の予測は，協調的フィルタリングによる評価の予測方法として利用されている．しかし，この方法において相関係数を求める際には，相関を求める利用者が共に評価しているアイテムしか利用することができないため，共に評価しているアイテムが非常に少ない場合には，協調的フィルタリングはうまく動作しないという問題がある．この問題に対する解決策の一つとして未評価のアイテムに対してデフォルト値を設定する方法があるが，どのような値を設定するかによって，予測は変化する．本稿では，評価の予測を行うアイテムに応じて適切なデフォルト値を設定する方法を提案する．映画評価データを用いた実験において，予測を行うアイテムによって適切なデフォルト値を設定することによって，精度の向上がみられた．

Study on Default Voting for Collaborative Filtering

Shigeki Muramatsu Naomi Inoue Keiichiro Hoashi Kazuo Hashimoto

KDD R&D Laboratories Inc.

Default voting is an extension to correlation based algorithm in collaborative filtering. Generally, when calculating Pearson correlation of users, as a weight, we use only votes in the intersection of the items which a pair of individuals have voted on. If we assume some defaul values, we can form the match over the union of voted items. But setting an appropriate default is not a trivial task because the default values change weights and predicted votes. In this paper, we decide a default value based on characteristics of an item. The result of experiments show that average absolute deviation of the predivted vote to the actual vote is decreased.

1 はじめに

情報フィルタリングは，膨大な情報の中から利用者にとって価値のある情報を選別するための技術である．情報フィルタリングには，内容そのものに基づいてフィルタリングを行う方法もあるが，その一方で，複数の利用者の好みの情報に基づいてフィルタリングを行う方法があ

り，協調的フィルタリングと呼ぶ[1][2][3][4]．典型的な協調的フィルタリングは，フィルタリングを行うアイテムの内容に関する情報そのものは利用せずに，他の利用者の利用状況や好みに基づいて情報フィルタリングを行う．

協調的フィルタリングのためのアルゴリズムには様々なものがある[5]が，一般的なアルゴリズムの一つに相関係数を利用するものがある[2]．しかし，利用者間の相関係数は共に評価を行っているアイテムを利用して求めるため，共に評価しているアイテムの数が非常に少ない場合には，適切な相関係数が求められないという問題がある．この問題に対する解決策の一つとして未評価のアイテムに対してデフォルト値を設定する方法があるが，設定した値により利用者間の類似度が変化し，結果として評価の予測値も変化するので，適切な値が設定されることが望まれる．

本稿では，設定するデフォルトの評価値を予測を行うアイテムによって変化させる方法を提案する．

2 従来の協調フィルタリング手法

2.1 相関係数を用いた協調的フィルタリング

協調的フィルタリングに要求される一般的な機能は，評価を蓄積したデータベースから特定の利用者の評価を予測することである．したがってデータベースは，アイテム j に対する利用者 i の評価 $v_{i,j}$ の集合によって構成される．ここで，I_i を利用者 i が評価したアイテムの集合とすれば，利用者 i の評価の平均は次のように定義できる．

$$\bar{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j} \qquad (1)$$

協調的フィルタリングによる評価の予測には，対象とする利用者の評価を他の利用者の評価の重み付け和から求めることができるという考えに基づいた方式があり，この場合，対象とする利用者(添字 a であらわす)のアイテム j に対する評価の予測 $p_{a,j}$ は，他の利用者の評価の重み付け和によって次のようにあらわされる．

$$p_{a,j} = \bar{v}_a + \kappa \sum_{i=1}^{n} w(a,i)(v_{i,j} - \bar{v}_i) \qquad (2)$$

ここで n は，重みが 0 でない利用者の数である．$w(a,i)$ は，利用者 a と i の距離や類似度をあらわす．また，κ は正規化係数であり，重みの絶対値の和の逆数を用いる．重みは，利用者間の好み等の類似度を何らかの尺度であらわしたものと考えることができるが，この尺度として用いられるものに相関係数がある．利用者 a と i の相関係数は次の式で計算される．

$$w(a,i) = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}} \qquad (3)$$

ここで j についての和は，利用者 a と i がともに評価を行っているアイテムについて求める．

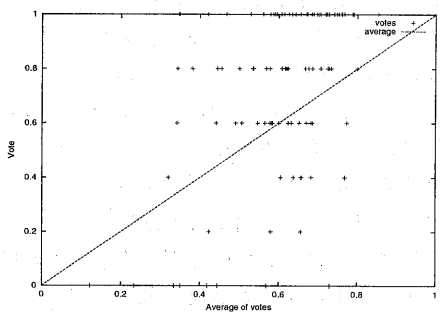
2.2 デフォルト投票

利用者間の類似度をあらわす尺度として相関係数を用いて利用者の評価を予測する方法の拡張の一つとしてデフォルト投票がある[5]．利用者 a と i の相関係数を求める際には，利用者 a と i がともに評価を行っているアイテム $(I_a \cap I_i)$ を利用するが，利用者 a または i の評価がほとんどない場合等にはともに評価を行っているアイテムがほとんどないため，適切な利用者間の類似度が求められず，協調的フィルタリングのアルゴリズムがうまく動作しないおそれがある．

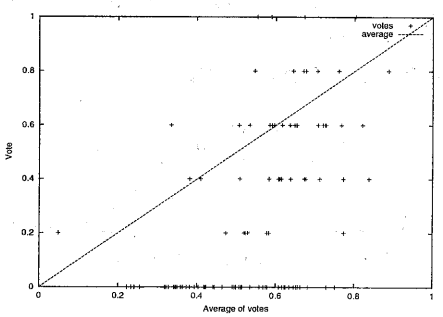
デフォルト投票は，評価の得られていないアイテムにある評価値を設定することにより，どちらか一方は評価したアイテム $(I_a \cup I_i)$ を相関係数を求める際に利用することで，この問題に対処する方法である．しかし，デフォルト投票の具体的な適用方法の報告は極めて少ない．

3 予測アイテムに応じたデフォルト値設定方式の提案

式 (2) は評価の予測値が，利用者が既に評価したアイテムへの評価の平均に他の利用者の評価に基づいた修正を加えることによって決定されることをあらわす．利用者に評価されたアイテムについて，利用者の評価の平均を横軸にアイテムへの評価を縦軸にとると，例えば，図 1 のように分布する．



評価値の平均が高い場合の例

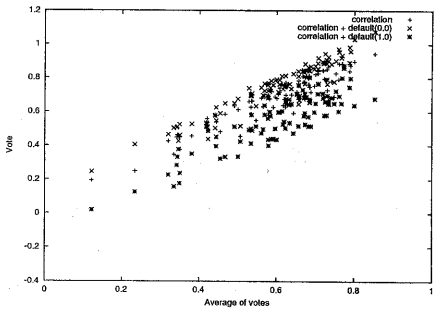


評価値の平均が低い場合の例

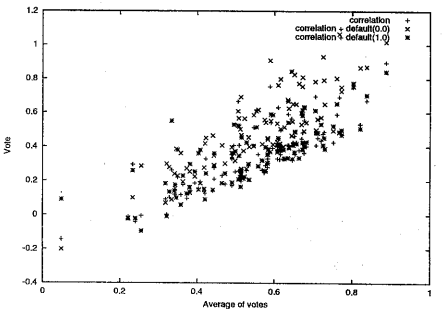
図 1: アイテムごとの利用者の評価の平均値とアイテムの評価

この評価の分布が，利用者の評価の平均をあらわす直線 $y = x$ よりも上の方に分布していれば他の利用者の評価に基づいた修正は評価を上げるように与えられるのが望ましく，反対に下の方に分布していれば評価を下げるような修正が適していると考えられる．図 1 の例であれば，最初のアイテムは利用者の評価の平均よりも高い評価が多く，2 番目のアイテムは利用者の評価の平均よりも低い評価が多くなっている．

実際に，いくつかの方法でデフォルトの評価値を設定して予測を行うと，図 2 のようになる．最初のアイテムは，デフォルトの評価値として



評価値の平均が高い場合の例



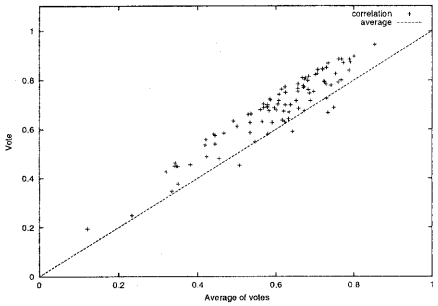
評価値の平均が低い場合の例

図 2: アイテムごとの利用者の評価の平均値と評価の予測値

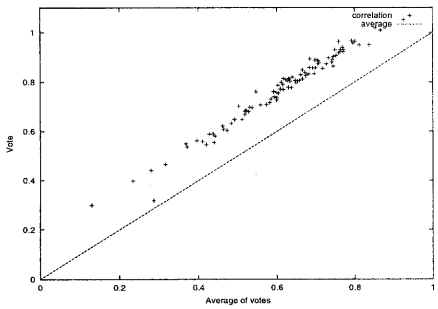
0.0 を設定した場合はデフォルト値を設定しない場合に比べ予測値は高くなり，1.0 を設定した場合は低い値になる．また，2 番目のアイテムでは，1.0 を設定した場合の予測値の変化は少ないが，0.0 を設定した場合には予測値は高くなる．このように，予測を行うアイテムの評価の平均が変わるとデフォルト値の設定による影響が変わる．

また，評価の平均が同じようなアイテムにおいても，評価を行っている利用者の数が多くなると，利用者の平均値と予測値の差が様々な値をとるようになる (図 3)．

このようにアイテムに応じて評価の分布や予測値の分布が異なっているので，予測を行うア



評価者の数が多い場合の例



評価者の数が少ない場合の例

図 3: アイテムごとの利用者の評価の平均値と評価の予測値

イテムに応じてデフォルトの設定値を変化させることによって，協調的フィルタリングの予測精度の向上が期待できる．

そこで，本稿では，予測を行うアイテムごとにデフォルトの評価値を決定するための以下のような方法を提案する．

- あらかじめ評価の蓄えられているデータベースをアイテムの評価の平均と評価者数に基づいて分割し，いくつかの方法でデフォルト値を設定し予測を行う．
- 分割したそれぞれの部分について最も精度がよくなる方法を求める．
- 予測の際には，予測を行うアイテムの評価の平均と評価者数に基づき最も精度が高い方法を選択し，デフォルト値の設定を行う．

4 実験と結果

4.1 実験用データセット

評価実験には，EachMovie[1] の評価データを利用した．EachMovie のデータは，Compaq Systems Research Center が18ヵ月間推薦システムを運用した結果得られた評価データである．利用者は，自分の観た映画について，0.0(最も悪い) から 1.0(最もよい) まで，0.2 刻みで 6 段階の評価を与えている．

実験のため，EachMovie のデータより 2 つ以上の映画を評価している利用者を選択した．その中から 5000 人を予測を行う利用者との相関係数を求めるための利用者として選択し，残りの利用者から実際に協調的フィルタリングによって評価の予測を行う利用者を選択した．

評価の予測は，利用者の評価データから予測するアイテムについての評価を除き，それ以外のその利用者の全ての評価を利用して行った．

条件を変化させた場合の予測の精度を評価するための尺度として予測値と実際の評価値との絶対分散を用いた．

各アイテムの絶対分散は，

$$S_j = \frac{1}{m_j} \sum_{a \in p_j} |p_{a,j} - v_{a,j}| \qquad (4)$$

のようになる．m_j はアイテム j について予測を行った利用者の数をあらわす．

4.2 結果

以下にデフォルト値の設定の方式を変化させ予測を行った場合の結果を示す．予測を行った利用者，およびアイテムは全ての方式について同じである．

結果中，CR は，デフォルト値を設定せずに相関係数を用いた場合をあらわす．AVE は，デフォルトの評価値として予測を行う利用者の評価の平均を用いた場合をあらわし，また，Def$n$ は，デフォルトの評価値として n を用いた場合である．

[1] http://www.research.compaq.com/SRC/eachmovie/

全てのアイテムについて同一の方法でデフォルトの評価値を設定し予測を行った場合の絶対分散の値を表1に示す．

続いて，評価の平均値と評価を行っている利用者の数に応じてデータベースを分割し，それぞれについて予測を行った．評価の平均値は，0.1刻みで10段階に分割した．また，評価者数については，評価者数とアイテム数の関係が，図4のようにあらわされるので，評価者数の範囲が指数的に増加するように設定した．

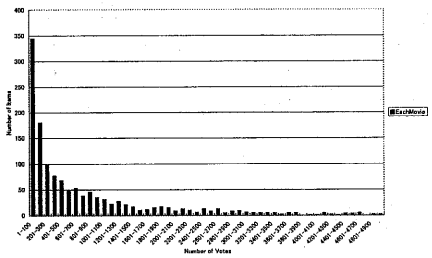


図 4: アイテムの評価者数の分布

分割されたそれぞれについて予測を行い，アイテム数が10以上の部分について最も精度の高かったものを選択すると，表2のようになる．

表2の条件で，デフォルトの値を設定した場合の結果を表3に示す．Bestは，各アイテムについて最も絶対分散の小さい場合を選択した場合である．また，相関係数を求める際に，デフォルト値を設定せずに利用できるアイテムの数の平均が10未満であったアイテムについてのみ同様に絶対分散を求めた場合を表4に示す．提案方式は，全ての場合に適用するためにはさらに精度を高める余地はあるが，相関係数を求めるのに利用できるアイテムが少ない場合には，効果があることが分かる．

表 3: 絶対分散の値

	CR	AVE	提案	Best
絶対分散	0.2095	0.2084	0.2073	0.2002

表 4: 絶対分散の値 (共に評価しているアイテムが少ない場合)

	CR	AVE	提案	Best
絶対分散	0.2051	0.2042	0.2005	0.1768

5 考察

一方で，予測と実際の評価の関係を個々のアイテムごとにみれば，予測を行うアイテムを評価している利用者が少なく相関係数を求めた利用者が少ない場合には，図5のように予測値が平均値から特定の方向にのみ変化し，結果として予測値が実際の評価値から大きくずれてしまうことがある．

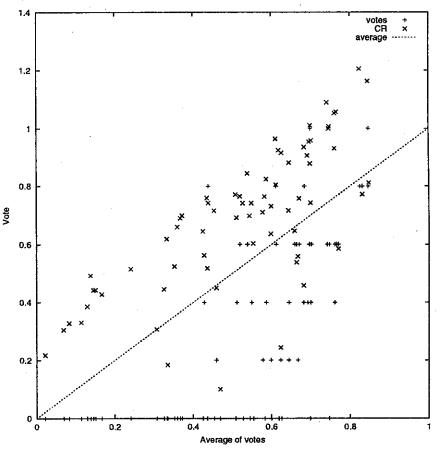


図 5: 相関を求めた利用者が少ない場合の例

また，利用者が多かった場合にも図6のような場合がある．この場合は，類似度の高い利用者と低い利用者に，評価の高い利用者，低い利用者が存在したため，結果として平均値からの修正項の絶対値が小さくなり，利用者の評価を予測できなかったものと考えられる．

以上のような点から，評価の平均値や評価者数といった全体的な特徴量に加え，個々の予測の分布についても考慮していく必要がある．

表 1: 絶対分散の値

	CR	AVE	Def0.0	Def0.2	Def0.4	Def0.6	Def0.8	Def1.0
絶対分散	0.2079	0.2067	0.2131	0.2109	0.2098	0.2111	0.2240	0.2548

表 2: デフォルト値の設定

	評価の平均値									
評価者数	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
0-50	-	Def0.2	Def0.2	CR	CR	Def0.0	CR	-	-	-
50-100	-	AVE	AVE	AVE	CR	Def0.6	Def0.6	-	-	-
100-200	-	-	CR	AVE	CR	AVE	CR	Def0.0	-	-
200-400	-	-	AVE	Def0.4	AVE	AVE	AVE	Def0.4	AVE	-
400-800	-	-	-	AVE	AVE	AVE	AVE	AVE	-	-
800-1600	-	-	AVE	Def0.6	Def0.4	AVE	AVE	AVE	AVE	-
1600-3200	-	-	AVE	Def0.6	AVE	AVE	AVE	AVE	-	-
3200-6400	-	-	-	-	Def0.8	AVE	AVE	AVE	-	-
6400-12800	-	-	-	-	-	Def0.4	AVE	-	-	-
12800-	-	-	-	-	-	-	AVE	Def0.2	-	-

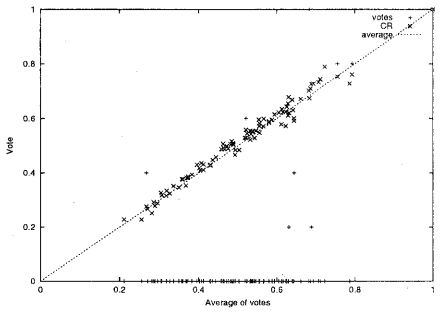


図 6: 相関を求めた利用者が多い場合の例

6 まとめ

他の利用者との類似度を相関係数を用いてあらわし評価の予測を行う協調的フィルタリングにおいて，相関係数を求めるために設定するデフォルトの評価値を予測を行うアイテムによって変化させる方法を提案した．評価を行うアイテムの評価の平均と評価者数に基づいて設定する方法では，予測精度に若干の改善は見られたが，最もよい場合と比較するとまだ精度を向上できる可能性があり，さらなる検討をする必要があると考えられる．

参考文献

[1] Paul Resnick and Hal R. Varian Recommender Systems, *Communications of the ACM 40(3)*, pp.56-58, 1997.

[2] Resnick, P., Iacovou, N. Sushak, M., Bergstrom, P., and Riedl, j. GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proceedings of the 1994 Computer Supported Collaborative Work Conference*, pp.175-186, 1994.

[3] Loren Terveen, Will Hill, Brian Amento, David McDonald and Josh Creter PHOAKS: A system for sharing recommendations. *Communications of the ACM 40(3)*, pp.59-62, 1997.

[4] James Rucker and Marcos J. Polanco Siteseer: Personalized navigation of the web. *Communications of the ACM 40(3)*, pp.73-75, 1997.

[5] John S. Breese, David Heckerman and Carl Kadie Empirical Analysis of Predictive Algorithm for Collaborative Filtering, *In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp.43-52, 1998.