Vol.2009-DBS-149 No.26
2009/11/21

MovieLens Dataにみる協調フィルタリングの失敗

高橋　徹[†1]　小林亜樹[†2]

Web 情報システムにおける情報推薦では，協調フィルタリング技術がしばしば用いられている．本稿では，その推薦精度が必ずしも高くない理由について分析する．分析には，MovieLens Data Set を用いる．協調フィルタリング技術の，類似した嗜好を持つユーザは，未知の商品などのアイテムにおいても同様の嗜好を示す，という仮定がどの程度正しいか分析する．その結果，この仮定が成り立たない場合が相当程度ある事を明らかにし，また，cold-start 問題として知られる，嗜好が蓄積されていない状況と似ている事を指摘する．その後，これらの問題に対処するための考察を進める．

Analysis of MovieLens data set about collaborative filtering failure

Toru Takahashi[†1] and Aki Kobayashi[†1]

Collaborative filtering techniques are often used to Information recommendation in the Web information system. In this paper, the reason why the recommendation precision is not always high is analyzed by using the Movie Lens data set. We think that collaborative filtering technique is based on a fundamental assumption. It is, if users had similar preference in an item set, it will be same as in other item sets. We show that the assumption is not approved in a respectable degree with similarity distribution histogram. Moreover, it is pointed out that the situation is similar to the cold-start situation, because of the same situation about user preference data is not available to collaborative filtering. Afterwards, it is considered to deal with these problems.

†1 工学院大学大学院工学研究科
Kogakuin University Graduate School of Engineering
†2 工学院大学工学部
Kogakuin University Faculty of Engineering

1. はじめに

Web サイトなどを通じた情報推薦システムの導入が活発である．情報推薦は，情報検索の一形態であると考えられるものの，ユーザの積極的な問い合わせを要さずにユーザ適意な情報を提供，あるいは提案する．ここで，推薦すべき情報には，ニュース記事や商品などが一般的であるが，これらを総称としてアイテムと呼ぶ．情報推薦技術では，ユーザの「嗜好」を抽出し，情報検索処理部に対する明示的な問い合わせを生成する手法が重要である．

ユーザの「嗜好」をアイテム自体の特性や属性に依拠するものとして記述しておき，合致するアイテムを提案するのが内容フィルタリングである．特定分野に限った情報推薦であれば，その分野のアイテムに特化した特性記述を行うことで，比較的精度良く推薦が可能となる．これは，情報検索技術の延長線上に位置する．

一方，多様な分野のアイテムを取り扱ったり，日々新しいアイテムが追加されるような状況では，それらの特性を体系化することは困難である．一方，可能な範囲による汎用的な特性記述を用いた内容フィルタリングでは，良好な推薦特性を得られないことは自明である．そのため，このような状況下では，協調フィルタリングが用いられるか，併用される．

協調フィルタリングは，ユーザの「嗜好」をアイテム自体の特性の「外」に求める．典型的には，システム内におけるユーザの利用履歴であり，端的にはアイテム毎の好みや評価（値）である．このとき，この「嗜好」情報自体の類似性を用いて，類似する「嗜好」を重視するようにしてアイテムの推薦度合い（推薦値）を計算する．協調フィルタリングは，アイテム自体の特性に依拠していないため，内容フィルタリングについて挙げた問題点とは無縁であり，導入が進んでいる理由もそこに求められる．

しかしながら，協調フィルタリグにおける「嗜好」情報自体が存在しなければ，推薦値を計算できない．これは，一般に cold-start 問題として知られている．また，「嗜好」の類似性がアイテムに依拠しないことが良好な推薦精度を保つための前提であるが，広範囲なアイテム間にあっては，これは直観に反する．協調フィルタリングベースの推薦手法において，必ずしも高い精度が得られない理由は，これらの問題に依る部分も大きいと考えられる．しかし，従来は推薦手法の提案と有効性の評価に終始する例が多く，失敗の原因に的を絞った研究は見ない．

そこで本稿では，協調フィルタリング研究を大きく進める端緒となった MovieLens Data Set を対象に，このような協調フィルタリングの失敗がどのようなデータの特性から発生しうるのかを分析し，その理由について明らかにする．さらに，協調フィルタリングが失敗す

る状況下において，内容フィルタリングの問題に回帰せず推薦を行うため，協調フィルタリング手法において用いている情報からどのような推薦を行うべきであるかについても議論する．

本稿の構成は次の通りである．まず，2 章で協調フィルタリングについて述べる．次に 3 章で解析に用いたデータセットである MovieLens Data Set と解析で用いるモデルについて説明し，4 章で 2 つのアイテム集合に見る相関の分布について解析する．5 章で想定される評価値とのずれについて解析し，最後の 6 章でまとめと今後の課題について述べる．

2. 協調フィルタリングの失敗

2.1 情報推薦における協調フィルタリング

一般に協調フィルタリングによる推薦とは，ユーザの利用履歴をユーザの嗜好としてとらえ，これをもとに嗜好が似たユーザ，もしくはアイテムを発見し，その情報を推薦に利用するという手法を指す．利用履歴により推薦を行うため，推薦するアイテムが Web ページや商品など，どのような場合においてもアイテム内容自体の解析が必要なく推薦を行える利点がある．

協調フィルタリングにはユーザベース方式とアイテムベース方式がある．本稿では前者の視点で解析を行うため，ユーザベース方式についてモデルを用いて説明する．ユーザが n 人，アイテムが m 個あり，ユーザ集合を $G = \{U_1, U_2, \ldots, U_n\}$，アイテム集合を $S = \{I_1, I_2, \ldots, I_m\}$ とする．ユーザベース方式では，推薦対象ユーザ U_t とユーザ U_u それぞれのアイテム I_l に対する評価値 $V(U_t, I_l)$，$V(U_u, I_l)$ により，ユーザ間の類似度 $P(U_t, U_u)$ を求め，推薦を行う．U_t に対する I_l の推薦度関数 $R(U_t, I_l)$ は，評価値のユーザ間での平均値の差を除いた本質に着目すれば，

$$R(U_t, I_l) = \frac{\sum_{U_u \in G} P(U_t, U_u) V(U_u, I_l)}{\sum_{U_u \in G} P(U_t, U_u)} \quad (1)$$

のように，類似度を重みとした加重平均として表せる．

一方，アイテムベース方式はアイテム間の類似度を用いて推薦を行う方式である．推薦手順は数学的にはユーザベース方式と同様であるものの，実装上の利点などから，採用例が多いとされる．

協調フィルタリングによる推薦の導入事例として，Amazon.com[⋆1]の「おすすめの商品」や、はてなアンテナ[⋆2]の「おとなりアンテナ」などがあるとされている．

2.2 問 題 点

協調フィルタリングを情報推薦に適用するにあたっては，その本質的な問題点としての cold-start 問題が指摘される[1)2)]．cold- start 問題は，推薦値の計算に不可欠な情報が存在しない点に起因し，推薦システム自体の初期立ち上げ期や，推薦対象となる人や商品などの個別の追加時などに発生する．推薦という観点自体からは他の何らかの情報から推薦することが可能であるにもかかわらず，このような推薦不可能な状況が発生する理由は，協調フィルタリングの枠組みで利用する情報が存在しないためである．したがって，この解決は，協調フィルタリングの枠組みの外にある情報をいかに汲み上げて推薦に活かしていくか，という視点で取り組まなければならない．

協調フィルタリングでは，ユーザの「嗜好」を用いて推薦値を計算する．この際，推薦システムが観測可能な「嗜好」情報において，被推薦ユーザ（アイテム）との類似度が高い他のユーザ（アイテム）は，被推薦対象アイテム（ユーザ）との類似度も高いはずである，という仮定に基づいている．これは，ユーザ（アイテム）間の類似度は，アイテム（ユーザ）に依らず高い相関があることを期待している．ただし，常に高い相関が必要なわけではなく，全体として高い相関の場合が存在し，相対的に低相関な場合が少なければ機能すると考えられる．この期待を仮定として単純化すると，ユーザ（アイテム）間類似度には高い相関が存在する，ということになる．すなわち，類似度間の高相関が維持されることが有用な推薦について望ましい．このような仮定が一定程度満たされていることは，これまでの協調フィルタリングベースの推薦手法や，それらを利用したと見られるシステム例が多数に上ることからうかがい知れる．

しかし，情報システム構築時のコンピュータ言語選択に関する嗜好が，好きなハンバーガーの種類に反映されると考えるには無理があるだろう．これは，同種の「嗜好」によってユーザがグループ化されうるアイテムのカテゴリの存在を意味し，逆に言えば，異なるカテゴリ間では協調フィルタリングベースの推薦が失敗する可能性が高いことを示唆する．実際に，精度の指標で協調フィルタリングが評価されるとき，それが常に高い数字とはならないことはよく知られている．

推薦にあたって有用な「嗜好」情報が，同種の嗜好を共有するようなアイテム（ユーザ）間からのみ得られるのだとすると，このような状況は広い意味での cold-start 状況である

⋆1 http://www.amazon.com

⋆2 http://a.hatena.ne.jp

とも言える．

2.3 本稿での解析の方向性

協調フィルタリングを情報推薦に適用した研究では，一定のデータセット（ユーザの嗜好情報）に対する，提案手法の有用性について議論することが主流で，データセット内固有の特性や，それと提案手法との相互作用などについて，データ内容の解析に基づく議論はあまり行われてこなかった．そこで本稿では，2.2 節で指摘したような，協調フィルタリングにおける精度低下の原因として疑いうる，広義の cold-start 状況が，実際にどの程度存在しているのか，を主題として解析を行う．

この観点からは多様な視点での解析が可能であるが，本稿ではユーザベースの協調フィルタリングの文脈における協調フィルタリングの失敗に絞って解析する．このとき，2.2 節の問題は，一定のカテゴリでは同種の「嗜好」を示す 2 ユーザが，他のカテゴリでは異なる「嗜好」を示すことが，どの程度存在するかという観測を行うことによって明らかになる．

ここで問題となるのは,「カテゴリ」というアイテムの集合をどのように定義づけるかである．先の議論では,「嗜好」が大きく異なる可能性の高い商品集合というニュアンスであった．しかし，このような商品集合を事前に定義することは難しい．古典的な「カテゴリ」に依拠する方法も考えられるが，この場合，古典的「カテゴリ」間に潜在する相関性などはないことを前提とすることになり，これは情報推薦の考え方と相容れない．一方で，実験的に「嗜好」の異なる商品集合を選択することは，そのような特殊な事例を選び出すことに繋がる．

そこで本稿では，4 章において，機械的に商品集合を分割し，その中でユーザ間の類似度間に存在する相関の状況について議論する．さらに，5 章では，協調フィルタリングが前提とする「嗜好」の類似の前提が利用できない状況下において，同様の「嗜好」情報から利用すべき情報について，議論の端緒を提供する．

3. データセットとモデル

3.1 MovieLens Data Set

本解析を行うにあたって用いる実データセットとして GroupLens[4] が提供している MovieLens Data Set を使用した（以下 MLDS）．MLDS はユーザが映画に対して行った評価のデータセットである．概要を表 1 に示す．MLDS は，ユーザ 943 人が映画 1682 件に対し，評価を 1 ～ 5 の 5 段階の評価で行い，総評価数は 100,000 件であるため，ユーザ一人当たりの平均評価数は 15 件となっている．

MLDS は協調フィルタリングに関する研究において用いられる代表的なデータセット[3] の 1 つである．新たなアルゴリズムの提案の文脈において MLDS を評価用のデータセットとして用いられている例も多く，下司ら[5]，神嶌ら[6]，木澤ら[7]，柳原ら[8] の研究で有用性の検証のため用いられている．しかし，MLDS 自体の解析を行った事例は知られていない．

表 1 MovieLens Data Set の概要
Table 1 Outline of MovieLens Data Set

評価対象 (映画)	1682(本)
評価人数	943(人)
評価数	100,000(件)
TimeStamp	100,000(件)

3.2 モ デ ル

本稿で用いる要素をモデル化する．データセット内のユーザ U_i は，評価人数が 943 人であるので,

$$U_i, (i = 1, 2, \ldots, 943) \tag{2}$$

また，U_i の中で推薦対象となるターゲットユーザを特に U_t として示す．データセット内すべてのアイテム I_j は

$$I_j, (j = 1, 2, \ldots, 1682) \tag{3}$$

とし，ユーザ U_i のアイテム I_j への評価は

$$V(U_i, I_j) = 1 \sim 5 \tag{4}$$

とする．

4. 2 つのアイテム集合にみる相関の分布

4.1 概　　要

ユーザベースの協調フィルタリングでは，推薦対象ユーザと評価値の類似度の高いユーザの情報が重視される．これが十分機能するためには，推薦対象ユーザと類似度の高いユーザとの間で，被推薦アイテムの評価値も十分類似する事が必要である．協調フィルタリングベースの推薦手法の興隆はこの仮定の正しさを一定程度示しているが，推薦精度の観点からは必ずしも正しくないと言える．

特にこの仮定の成立が危ぶまれるのは，類似度を計算するアイテム集合の特性と，被推薦アイテム（集合）の特性とが大きく異なる場合である．例示としては，数学の参考書と映画のように大きく異なる分野（カテゴリ）に属する商品などである．

そこで本章では，類似度計算アイテム集合による類似度が被推薦アイテム集合の類似度に

対してどの程度類似しているのかを明らかにする事を目的に，MLDS を複数のアイテム集合に分割し，各アイテム集合毎に評価されたデータを対象とした類似度を計算し，その分布を調査する．本来は，アイテム集合の分割は映画の種類に依存するような分割法が望ましいと言える．しかしここでは，機械的な分割であっても，各集合内での類似分布が必ずしも高い相関を示す訳ではない事を示す事で，嗜好が大きく異なると考えられるような異なるカテゴリ間での推薦に対する協調フィルタリングの限界について指摘する．

4.2 アイテム集合の分割

異なるアイテム集合間における，同一組のユーザ間の類似度の分布を知る事を目的とする．そこで，データセット内の全てのアイテムの集合 $S = \{I_1, I_2, \ldots, I_{1682}\}$ を

$$S = S_a \cup S_b, S_a \cap S_b = \emptyset \tag{5}$$

のように，アイテム集合 S を 2 つの互いに排他なアイテム集合 S_a と S_b に分ける．

4.3 解 析 手 順

本解析では，機械的に分けたアイテム集合 S_a，S_b において，任意のユーザ間の類似度 P_a，P_b をピアソンの積率相関係数で求め，その組み合わせの分布を調べる．図 1 に概念図を示す．あるユーザ U_t と U_u 間の評価値の類似度 P_a，P_b を，アイテム集合 S_a，S_b それぞれについて求める．協調フィルタリングが良好な特性を示すには，これらの類似度 P_a，P_b は一致する事が望ましい．そこで，これらの関係を P_a を水平方向，P_b を垂直方向の座標とした度数分布表にプロットする．なお，本解析では P_a，P_b の度数分布を観測する事を目的とするため，P_a，P_b 間の相関係数を示す事はしない．

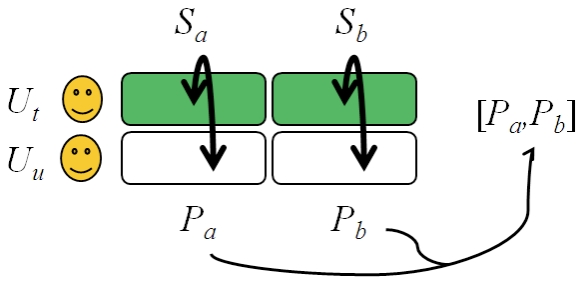

図 1 解析対象の類似度組
Fig. 1 Similarity tuple of analisys

具体的な手順は次の通りである．

(1) S を分割するにあたり，一方（S_a）の大きさを，100，500，800，1000 とする．
(2) 定めた S_a の大きさに基づき，機械的に 2 つの互いに排他な集合，S_a, S_b に分割する．
(3) 各 S_a, S_b における評価値の類似度 (P_a, P_b) を任意の 2 ユーザ U_i，$U_j (i \neq j)$ について求める．
(4) 類似度 P_a, P_b の分布を S_a の大きさごとに集計する．

4.4 ピアソンの積率相関係数

類似度関数として，ピアソンの積率相関係数を用いる．ピアソンの積率相関係数は，2 つのデータ列を比較して類似度の度合を求める関数であり，ユーザベースの協調フィルタリングでは，ユーザごとの評価値列が用いられる．これは，多くの協調フィルタリングのアルゴリズムにおいても利用されてきた．

2 組の数値からなるデータ列 $(x, y) = (x_i, y_i)(i = 1, 2, ..., n)$ が与えられたとき，ピアソンの積率相関係数 P は，

$$P = \frac{\sum_{i=1}^{n}(x_i - x_{ave})(y_i - y_{ave})}{\sqrt{\sum_{n=1}^{n}(x_i - x_{ave})^2}\sqrt{\sum_{n=1}^{n}(y_i - y_{ave})^2}} \tag{6}$$

と表される．ここで x_{ave} と y_{ave} はそれぞれのデータ列の相加平均を表す．

P は $-1 \leq r \leq 1$ であり，

- 0 に近い：二変数は無相関
- 1 に近い：正の相関
- -1 に近い：負の相関

ということをそれぞれ示す．P の絶対値が大きい場合に，推薦元データとして大いに利用できるとされる．

4.5 結果と考察

図 2 ～ 図 5 にそれぞれ，S_a を，100，500，800，1000 とした時の結果を示す．横軸が P_a，縦軸が P_b であり，階級幅は 0.1 である．ただし，階級値 1 は，類似度 P_a または P_b が 1 の場合のみを計数している．階級値に属する (P_a, P_b) の組の数を各階級を示す領域に，数が増えるにつれ濃くなるよう色分けして示している．

S_a を，100，500，800，1000 と変化させた場合すべてに共通して図の中心の辺りは濃く，P_a，P_b が共に 0 に近い組み合わせ，つまり S_a，S_b で互いにユーザ間の相関が低い組み合わせが多いということがわかる．これらのデータ組は協調フィルタリングにおいて情報が重視されない組み合わせである．

では，協調フィルタリングを用いた推薦で重視される組み合わせとは，どのような組み合

Vol.2009-DBS-149 No.26
2009/11/21

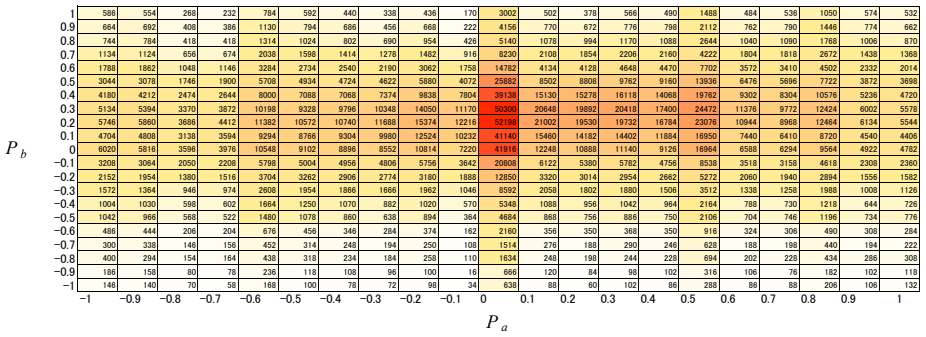

図 2 アイテム集合に対する類似度組の度数分布（$S_a = 100$）
Fig. 2 Histogram of similarity tuple for two item sets ($S_a = 100$)

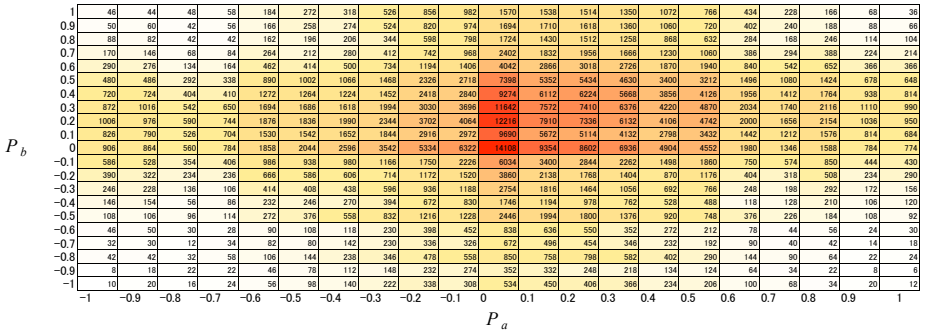

図 3 アイテム集合に対する類似度組の度数分布（$S_a = 500$）
Fig. 3 Histogram of similarity tuple for two item sets ($S_a = 500$)

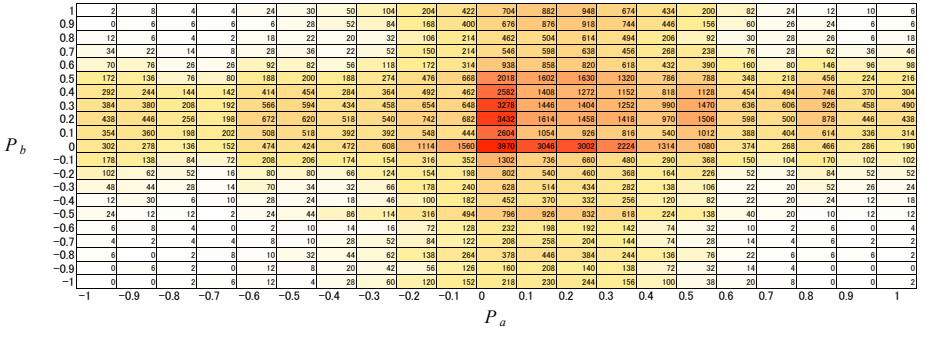

図 4 アイテム集合に対する類似度組の度数分布（$S_a = 800$）
Fig. 4 Histogram of similarity tuple for two item sets ($S_a = 800$)

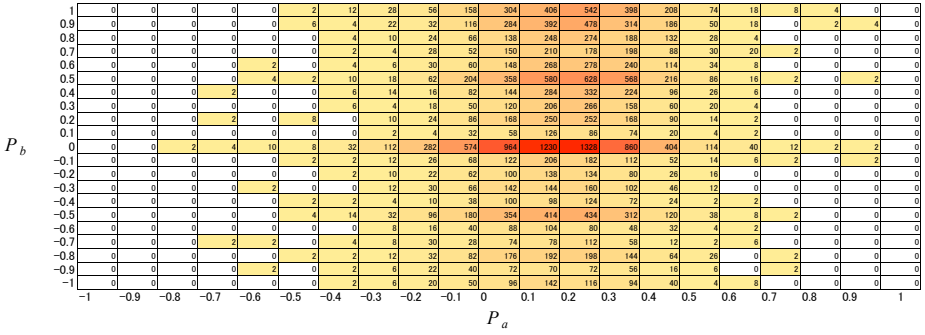

図 5 アイテム集合に対する類似度組の度数分布（$S_a = 1000$）
Fig. 5 Histogram of similarity tuple for two item sets ($S_a = 1000$)

わせなのだろうか．

図 6 に推薦値の計算がどのように行われるかの概念図を示す．図 6 ではアイテム集合 S_a に対する評価値はユーザ U_t，U_u 共に存在しており，一方，S_b に対する評価値は U_u では存在するが，U_t では存在していない．ここでは，U_t の S_b が被推薦対象，すなわち，U_t による S_b の評価値予測として推薦値の算出が行われる．このような状況下においては，P_a は求められるが，P_b は求める事ができない．この場合，協調フィルタリングによる推薦を行うと，P_a の値が高いユーザ情報を重視し S_b に対する推薦を行う．ここで協調フィルタリングが良好な特性を示すのは，P_a と U_t の S_b に対する真の評価値がわかったときの P_b とが，共に 1 に近い類似度である場合である．また，共に -1 に近い類似度である場合も相関が強いという事により，協調フィルタリングが良好な特性を示すと考えられる．このような組み合わせが多く存在すれば，協調フィルタリングによる推薦は精度良く機能する事が予想される．では，これらは全体の評価の組み合わせに対し，どの程度存在しているだろうか．

図 2 ～ 図 5 において P_a，P_b がそろって 1 に近いか，または-1 に近いのは，各図においておおむね右上と左下の範囲となる．

Vol.2009-DBS-149 No.26
2009/11/21

ここで，類似度間の相関関係の理解を助けるために，図 2 を太線で 9 つに分割した図 7 に注目する．P_a，P_b が共に強い正の相関，負の相関を示している組み合わせは，図中の 3 と 7 の領域であり，これらの相関度数は全体に対し 12%しかない．一方，重視されるが失敗する組み合わせが，図中の 1，9 の領域に相当する．この領域に含まれるデータ組の相対頻度は，全体に対して 6.6%であった．この，失敗する 6.6%の組み合わせの存在と，良好な特性を示す組み合わせが 12%しか存在していない事から協調フィルタリングの推薦精度が低いという問題が生じているものと考えられる．

本稿では，映画というある程度カテゴリのそろったアイテム集合を分割して解析を行ったものの，このような結果が得られた．この事より，さらに異なるカテゴリに属するアイテム間における場合は，この結果以上に色濃く影響を及ぼす考えられる．

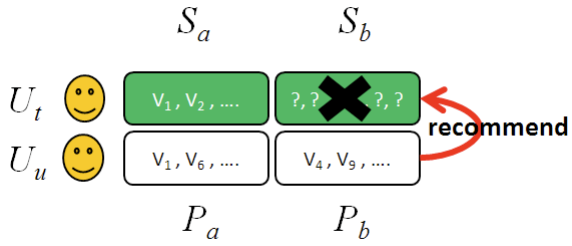

図 6 推薦の概念図
Fig. 6 Outline of recommendation

5. 想定される評価とのずれ

5.1 概　　要

5.6 節で述べたように，協調フィルタリングを用いた推薦において，活用され，また，有効に機能する情報の組み合わせは，全体の 12%程度である．そのため，推薦精度の低下という問題が発生するものと考えられる．では精度を向上するにはどのようなユーザ情報を用いることが好ましいのであろうか．

類似度の高いユーザ情報のみを用いたときに，推薦に失敗するような状況下では必然的にそのようなユーザ情報のみを使う事は得策ではない．ユーザ情報の活用が難しいという意味では，そもそも，そのような情報の無い cold-start 状況に等しい．そのため，既に情報の蓄積されている他の情報源の情報を利用する戦略[8] が考えられる．しかし，被推薦アイテム

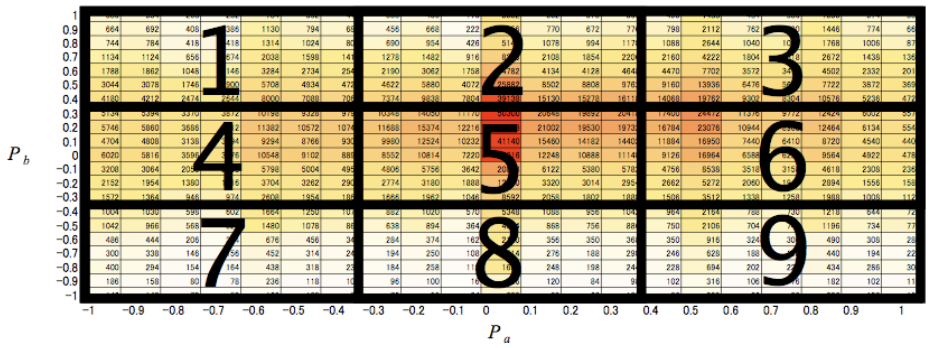

図 7 類似度によるエリア分割（$S_a = 100$）
Fig. 7 Area division by similarities($S_a = 100$)

集合に類似の情報が無ければ，状況に改善は見込めない．

また，内容フィルタリングによる研究も報告されている[9]．しかし，質の異なる情報であるため有効に機能するものの，協調フィルタリングの利点を減殺してしまう面を持つ．そこで本稿では協調フィルタリングと同一のユーザ情報，すなわち評価値列の利用可能性を探る．

質の高いユーザ情報とは，信頼性のあるユーザの評価であると考えられ，これを用いる事で推薦精度の向上が期待できる．ここで信頼性のあるユーザとは，4 章のように異なるアイテム集合が存在する状況において，アイテム集合ごとに総評価数が多いユーザの情報であるという仮定とおく．なぜなら，アイテム集合内での総評価数が多いユーザとは，その集合内において，アイテムに評価数の少ないユーザと比べて詳しいユーザである．そのため，評価の信頼性も高いと考える．そこで，評価数の多少が評価の信頼性に及ぼす影響について解析する．

5.2 ユーザ集合

ユーザ U_i の総評価数を $N_v(U_i)$ とする．この評価数により，評価数の多いユーザを上位ユーザ集合，評価数の少ないユーザを下位ユーザ集合と呼び，それ以外のユーザを中位ユーザ集合とする．上位ユーザ集合，下位ユーザ集合はデータセット内の総ユーザ数 943 人の 10%にあたる 94 人になるようにした．94 人目の総評価数と同じ総評価数のユーザは同じユーザ集合になるようにした．具体的には，上位ユーザ集合は総評価数が 245〜737 個の 94 人とし，下位ユーザ集合は評価数が 20〜23 個の 99 人とした．また，中位ユーザ集合は評価数が 24〜242 の 750 人とした．表 2 には各ユーザ集合の定義を一覧で示した．

表 2 ユーザ集合
Table 2 User sets

ユーザ集合	定義	人数
上位ユーザ集合	$U_h = \{U_i \mid N_v(U_i) \geq 245\}$	94
中位ユーザ集合	$U_m = \{U_i \mid 23 < N_v(U_i) < 245\}$	750
下位ユーザ集合	$U_p = \{U_i \mid N_v(U_i) \leq 23\}$	99

各ユーザ集合毎の評価値分布を図 8 に示す．これは各ユーザ集合内において $V(U_i, I_j)$ の値がどのように分布するのかヒストグラムにより表している．いずれのユーザ集合にあっても，評価値の与え方に大きな偏りは見られず，評価値 4 を与える映画が最も多く，次いで 3，5 の評価である．したがって，各ユーザ集合ともに傾向の似た評価をしているという事がわかる．

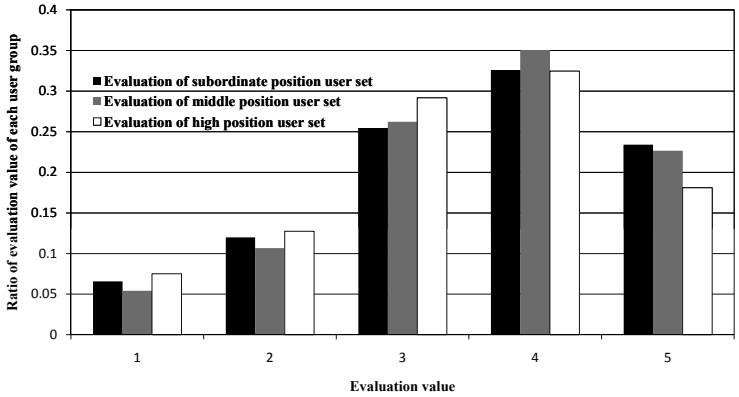

図 8 各ユーザ集合の評価値相対度数分布
Fig. 8 Relative histogram of evaluated values for each user set

5.3 解　　析

あるアイテムにおいて，データセット内の 80%のユーザで構成される中位ユーザ集合の評価の平均を推薦されるべき評価と仮定する．このとき，上位ユーザ集合，下位ユーザ集合それぞれの評価の平均と中位ユーザ集合の平均の差を求め，推薦されるべき評価に対し，どの程度ずれが生じるのかを観察する．本解析において，ずれの正負は重要でないため絶対値を取る事とする．I_j での上位，中位，下位，それぞれのユーザ集合の評価の平均は V_{hj}，V_{mj}，$V_{pj}(j = 1, 2, \ldots, 1682)$ となる．また，上位ユーザ集合，下位ユーザ集合それぞれの評価のずれは

$$G_{hj} = |V_{hj} - V_{mj}| \quad (7)$$

$$G_{pj} = |V_{pj} - V_{mj}| \quad (8)$$

である．

横軸をずれの絶対値，縦軸を上位，下位，双方のユーザ集合内での相対頻度としたヒストグラムを図 9 に示す．折れ線は上位ユーザ集合と下位ユーザ集合の相対頻度の差である．線が正方向に多い場合は上位ユーザ集合の評価の方が多く，負の場合は下位ユーザ集合の評価の方が多い事を示す．

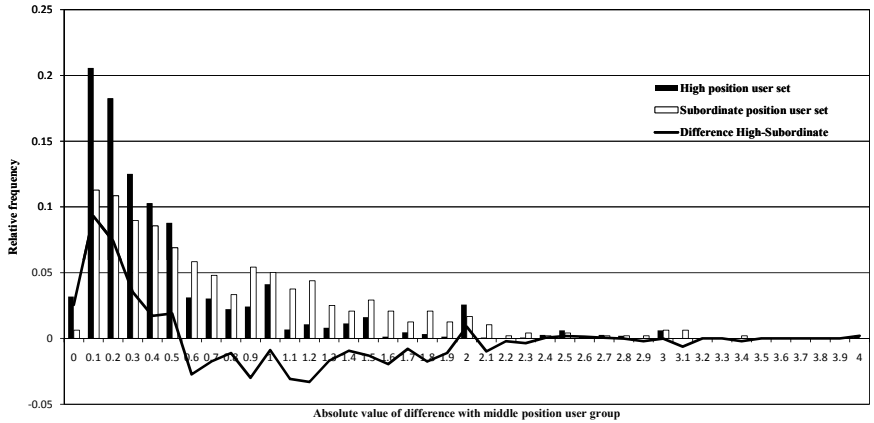

図 9 評価値のずれの相対頻度
Fig. 9 Relative frequency of evaluation difference

5.4 考　　察

本解析において，中位ユーザ集合は与えられるべき評価をしているユーザ集合と仮定した．よって横軸の評価平均の差の絶対値は推薦のずれの度合いを表す．

上位ユーザ集合は横軸 0 ～ 0.5 の範囲に約 74%の評価が存在している．一方，下位ユー

ザ集合では同範囲に評価が約 47%に留まっている．また，横軸の値 0.6 以上で下位ユーザ集合の評価が上位ユーザ集合より多い状況であり，相対的に評価値のぶれが大きい．評価数が多いユーザ集合である上位ユーザ集合の評価は誤差が少なく比較的安定した評価であるのに対し，評価数の少ないユーザ集合である下位ユーザ集合は評価にばらつきが見られる．

本解析では，ユーザ集合ごとの評価値の平均のみを対象としたため，上位ユーザ集合の評価数の方が下位ユーザ集合の評価数より多い事から，単に大数の法則に従い，このような結果を見せている事も考えられる．したがって，推薦の文脈においては個々のユーザ単位への推薦値の「ぶれ」量として，各ユーザ集合の評価を行うべきであり，今後の課題といえる．

しかし，同数のユーザ情報を用いるなどの単純な実装や，また，計算速度の制約などを考慮する事も必要である．その場合，評価数の多いユーザと少ないユーザでは，多いユーザの評価の方が信用をおけるという見方ができる．したがって，評価数が多いユーザの情報に基づく推薦を行う事が好ましいと考えられる．

5.5 cold-start 問題の解決方針

新規ユーザを被推薦対象とした場合，協調フィルタリングでは履歴による嗜好算出を行い推薦に用いるため，利用可能な情報は限られる．このような状況は cold-start 問題として知られる．このような状況下では，被推薦対象ユーザの嗜好情報が抽出困難であるが，推薦されたいアイテムのカテゴリ，例えば音楽 CD であればそのカテゴリ内での評価数の多いユーザの情報を重視して推薦を行う事で，善処できるものと考える．

6. 終わりに

本稿では，協調フィルタリングによる推薦における精度低下要因について，MLDS を例に取り解析を行った．筆者らは，精度低下の要因として，ユーザ間における嗜好情報の類似度が大きく異なる，すなわち，ユーザ毎の嗜好傾向が大きく異なるようなアイテム集合が存在するのではないかと考えた．MLDS のアイテム集合を機械的に 2 分割し，各集合内でのユーザ間類似度が示す相関傾向について，度数分布をグラフ化し，その特徴について論じた．

その結果，協調フィルタリングによる推薦に効果的に用いられるデータ組はわずかであり，逆に精度を引き下げる要因となるデータ組の割合がその半分に上ることが明らかになった．これは，協調フィルタリングによる推薦が前提としている，ユーザ間における嗜好情報の類似度は，アイテム集合に（ほとんど）依存しない，という仮定が成立していない場合が相当数に上ることを意味する．

このような場合には，被推薦対象ユーザの既知の嗜好情報を用いた協調フィルタリングは，推薦精度の観点からは逆効果となる．言い換えれば，協調フィルタリングとして利用可能な情報がない状況であると言え，その意味では，cold-start 問題と状況として似ていると言える．cold-start 問題に対処するため，他の情報源の嗜好情報を用いるなどの手法も報告[8] されているが，アルゴリズムとしては協調フィルタリングであるため，アイテム集合間の評価傾向が大きく異なる場合には，推薦精度の低下が予想される．

内容フィルタリング等，他の手法との組み合わせの有効性は徐々に報告されつつある状況であるが，本稿では，協調フィルタリングで用いるのと同一の情報源のみから，利用可能性のある情報について考察した．その結果，嗜好情報の蓄積の多いユーザの情報を，実用上の信頼度が高いと見なすことが有効ではないか，との示唆を得た．既に筆者らは，その方針に基づいた手法について提案しており，手法や有効性の評価などの改善は今後の課題である．

謝　辞

本研究の一部は，科学研究費補助金 基盤研究 (A)20240072 による．ここに記して謝意を表します．

参 考 文 献

1) 土方嘉徳，神嶌敏弘，市川裕介，河合由起子，村上知子，小野智弘，本村陽一，麻生英樹，乾孝司，奥村学，金山博，“利用者の好みをとらえ活かす-嗜好抽出技術の最前線-，” 情処学誌，vol.48, no.9，pp.955-1007，2007.
2) 川前徳章，鈴木英明，水野修，“情報・知識共有を支援する協調フィルタリングに関する研究-協調フィルタリングの適用箇所拡大にむけた方法の提案-，” 信学技報，PRMU103(389)，pp.37-42, 2003.
3) Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, John, T. Riedl “ Evaluating collaborative filtering recommender systems, ” ACM Transactions on Information Systems, vol.22, pp.5-53, 2004.
4) GroupLens, “http://grouplens.org”
5) 下司義寛，廣川佐千男，“概念グラフを使った推薦，” 電子情報通信学会第 9 回 Web インテリジェンスとインタラクション研究会，2007.
6) 神嶌敏弘，赤穂昭太郎，“転移学習を利用した集団協調フィルタリング，” 第 32 回人工知能学会全国大会，2009.
7) 木澤寛厚，磯崎邦隆，菊池浩明，“秘匿積集合プロトコルを利用したプライバシ協調フィルタリングの提案，”2009 年暗号と情報セキュリティシンポジウム，2009.
8) 柳原正，帆足啓一郎，松本一則，菅谷史昭，“次元圧縮を用いたクロスメディアレコメンデーション方式の提案，”FIT，一般講演論文集第 2 分冊，pp.83-86，2007.
9) Deok-Hwan KIM, “Image Recommendation Algorithm Using Feature-Based Collaborative Filtering, ” IEICE trans. Vol.E92-D, No.3, pp.413-421, 2009.

研究報告 2009-DBS-149-26

正誤表

6 ページ目

5.1 概要の項

×5.6 節で述べたように　→　○4.5 節で述べたように