

## 協調フィルタリングにおけるユーザの選択及び 嗜好抽出の手法の提案

川前徳章 向垣内岳弥 花木三良

NTT情報流通プラットフォーム研究所  
〒180-8585 東京都武蔵野市緑町 3-9-11

E-mail: {kawamae.noriaki, mukaigaito.takeya,hanaki.miyoshi}@lab.ntt.co.jp

あらまし 本研究は、協調フィルタリングの予測精度の向上と計算量の低減を実現する新しい手法を提案する。協調フィルタリングの有効性はその予測精度によって評価され、予測精度は嗜好が類似したユーザ間の発見方法に依存する。従来手法は類似したユーザをユーザの評価の類似度を用いたが、提案手法はユーザを選択し、その選択したユーザの評価から抽出した嗜好の類似度を用いる。本研究はこのユーザの嗜好を抽出するためのアイテムに対する重み付けの手法、ユーザ選択の手法、潜在嗜好モデルの導入とユーザの評価の予測式を提案する。これらを用いて実験を行った結果、提案手法の有効性を予測精度、計算量の面で確認することができた。

キーワード 協調フィルタリング 知識共有 潜在嗜好 履歴分析

## Extracting User's Preference and Selecting Users in Collaborative Filtering

KAWAMAE noriaki, MUKAIGAITO Takeya, HANAKI Miyoshi

NTT Information Sharing Platform Laboratories

3-9-11, Midori-cho Musashino-shi Tokyo 180-8585 Japan

E-mail: {kawamae.noriaki, mukaigaito.takeya,hanaki.miyoshi}@lab.ntt.co.jp

**Abstract** This paper proposes a novel method in collaborative filtering in terms of the accuracy of prediction and calculation cost. The effectiveness of collaborative filtering can be assessed by its accuracy of prediction, depending on the way to seek similar users in preference. Previous works seek these users by the similarity between the users' rating. But our method selects users and measures the similarity by the users' preference. To extract these users' preference, we propose a weighting method for item, introduction of users' hidden preference model and a formula to predict users' rating. Besides these, we propose the formula to select users from all users quantitatively for extracting user's preference. Through the experiments, we can confirm the accuracy of prediction based on extracted user's preference and reduction of calculation cost by selecting users.

**Keyword** Collaborative Filtering, Knowledge Sharing, User's Hidden Preference, History Analysis

### 1. はじめに

#### 1.1. 協調フィルタリングによる情報収集の効率化

協調フィルタリングは、個々のユーザの持つ知識や情報を共有することで、類似したユーザを発見し、そのユーザ間で知識や情報を提供し、ユーザの情報収集行動を効率化する手法である。協調フィルタリングが行う代表的なサービスにはナレッジナビゲーション、リコメンデーションがある。

ナレッジナビゲーションを会社内の業務で利

用すると、ユーザは同じ問題意識を共有する現在あるいは過去のユーザが利用した文書、資料やメールの利用履歴を利用して、その問題解決に必要な文書、資料やメールなどを効率的に見つけることを支援できるようになる。その結果、会社内での知識共有が進み、業務が効率化される。リコメンデーションをECサイトで利用すると、ECサイトは過去に購入した書籍やCD、DVDなどの商品の購入履歴を利用して、そのユーザが購入する可能性のある商品を予測して提示することが可能

となる。その結果、EC サイトはユーザの購買機会の拡大やロイヤリティの増大による収益の増大だけでなく広告やプロモーションにかかる費用の低減も望めることができる。従って、今後も協調フィルタリングは会社内の業務、EC サイトを始め、多くの分野での利用が期待される。

### 1.2. 協調フィルタリングの課題

協調フィルタリングが、今後多くの分野で利用される為には、その予測精度の向上と計算量の低減を実現する必要がある。予測精度はユーザの評価に対する協調フィルタリングの予測の精度であり、計算量はその予測を算出するのに要する時間である。予測精度は類似したユーザの発見方法に依存するので、適切な発見方法が必要である。計算量については、従来は全ユーザの評価データを用いている為にユーザの数の増加に伴って計算量が増加する問題がある。

### 1.3. ユーザ選択と嗜好の抽出による問題解決

本研究の目的は協調フィルタリングの予測精度の向上と計算量を低減することである。提案手法は、予測精度の向上の為に、類似したユーザを、ユーザの嗜好の類似度により発見し、計算量の低減の為に全ユーザから計算に用いるユーザの取捨選択を行うことで問題を解決する。

## 2. 従来研究とその問題点

代表的な協調フィルタリングとして知られている手法に GroupLens[9] 及び Ringo[11] がある。これらの手法は、あるユーザのアイテムに対する評価の予測を他ユーザの評価データを用いて行う。評価の予測にはユーザ間の類似度を必要とし、いくつかの類似度の計算手法が提案されている [1] [9] [6]。Ringo[11]、The Bellcore Video Recommender[7]はユーザ間の類似度をピアソンの積率相関を用いて計算している。Breese ら[1]は Each Movie[12]において類似度の中でピアソンの積率相関を用いたものが最も精度が高いことを報告している。これらの手法は共通して、類似したユーザの発見に、全ユーザのデータを用いることと、評価の類似度を用いている問題点がある。

### 3. ユーザの潜在嗜好モデルとユーザの選択

本研究の目的は協調フィルタリングの予測精度の向上と計算量を低減することである。予測精度はこの類似したユーザの発見方法に依存する。

従来手法は類似したユーザを次のように発見

していた。あるユーザ  $i$  がアイテム「Enya」、「Deep Forest」に対して「好き」という評価を数値化し、この数値を用いることでユーザの類似度を定義する。嗜好の類似したユーザを発見はこの類似度を用いて行っていた。従ってユーザのアイテムに対する評価を用いて類似度を測定する場合、ユーザの類似度はユーザの評価の類似度になり、発見されるユーザは評価の類似したユーザになる。

提案手法は類似したユーザの発見に、ユーザの評価データの背後にある嗜好を抽出し、その嗜好の類似度を用いる。ユーザの嗜好はユーザのアイテムに対する評価から抽出する。その為にアイテム毎の重み付けと潜在嗜好モデルを提案する。例えば、先の例では、アイテムの「Enya」、「Deep Forest」が嗜好の「ヒーリング系の音楽」に相当する。この嗜好とユーザ  $i$  の関係を数値化し、提案手法はこの数値を用いて嗜好の類似したユーザを発見する。その結果、発見されるユーザは嗜好の類似したユーザになる。ユーザの発見方法が評価でなく、抽出した嗜好を用いることで予測精度の向上が実現できる。図1にユーザ、ユーザの嗜好とアイテムの関係を示す。図1の上はユーザの評価データに従来手法を用いたもの、下は同じデータに、提案手法を用いた時に、抽出されるユーザ、嗜好とアイテムの関係を示す。

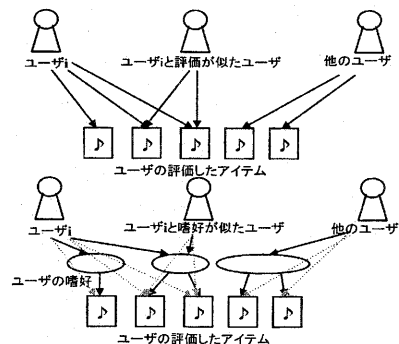


図1 ユーザ  $i$  に対する嗜好とコンテンツの関係

従来手法は類似したユーザの発見に全ユーザの評価データを用いていた。従って、ユーザの数の増加に伴い計算量も増加する。提案手法はユーザの嗜好の抽出に必要なユーザを取捨選択することでこの問題を回避する。

提案手法はユーザの嗜好を抽出する為に次の四段階から構成される。

1. アイテム毎に全ユーザの評価に基づいた重み付け
  2. 全ユーザから嗜好の抽出に必要なユーザを選択
  3. ユーザの嗜好をユーザの潜在嗜好モデルによって抽出
  4. ユーザの評価を2で抽出したユーザの嗜好に基づいたユーザ間の類似度により予測
- 1については3.1、2については3.2、3については3.3、4については3.4において詳細を述べる。

### 3.1. アイテム毎の重み付けとその選択手法

ここでは全ユーザのアイテムに対する評価を利用してアイテム毎の重みを定義する。アイテム毎に重みを付けるのは、ユーザの評価したアイテムには、そのユーザの嗜好を抽出するために役立つものとそうでないものがあると考えられるためである。どのユーザからも同様に評価されるアイテムはユーザの嗜好を区別するのに重要でないと考えられる。逆にユーザによって評価の異なるアイテムはユーザの嗜好を区別するのに重要と考えられる。そこで、本研究は、アイテム毎の重みにユーザの評価の異なり具合を利用する為に以下の式によって定義する。

$$w_{ij} = r_{ij} \cdot g_j = r_{ij} \cdot \left( 1 - \left( -\frac{1}{\log_2 n_j} \sum_{i \in n_j} rr_{ij} \log_2 rr_{ij} \right) \right) \quad (1)$$

- $w_{ij}$  : ユーザ  $i$  のアイテム  $j$  に対する重み
- $r_{ij}$  : ユーザ  $i$  のアイテム  $j$  に対する評価
- $g_j$  : アイテム  $j$  の重み
- $n_j$  : アイテム  $j$  を評価したユーザの数
- $rr_{ij}$  : ユーザ  $i$  のアイテム  $j$  に対する相対評価

$$rr_{ij} = \frac{r_{ij}}{\sum_{i \in n_j} r_{ij}} \quad (2)$$

我々はアイテム毎の重み  $g_j$  にエントロピーを用いる。エントロピーは熱力学で定義されたものであるが、事象の情報の量を定義することもできる。この場合、事象はアイテム、情報の量はそのアイテムに対するユーザの評価に相当する。従って、エントロピーを用いることで、各アイテムにユーザの評価を用いて重みを付けられる。提案手法はこの重み  $g_j$  をアイテム毎に全ユーザに対して付けるので、同一アイテムは全ユーザで同一の重みを持つ。エントロピーは各事象が一様分布に従っ

て生起する時に最大となり、生起確率が偏ると小さくなる。この式ではアイテム  $j$  を評価したユーザの数  $n_j$  で正規化を行った後に1から引く。その結果、アイテムが全ユーザに同じ評価をされた場合、そのアイテムの重みは0になり、ユーザによって評価が異なる場合はその重みが1に近くなる。この重み  $g_j$  にユーザ  $i$  のアイテム  $j$  に対する評価  $r_{ij}$  をかけてユーザ  $i$  のアイテム  $j$  に対する重み  $w_{ij}$  を算出する。

### 3.2. ユーザの選択手法

ここでは全ユーザからユーザの嗜好の抽出に必要なユーザを選択する。ユーザの嗜好の抽出は既存手法と同様にユーザの評価データを出発点として用いる。しかし、この評価データの中には嗜好の抽出に役立つものとそうでないものが存在する。本研究で抽出する嗜好は個人の絶対的な嗜好でなく、ユーザ全体を区別する相対的な必要嗜好である。従って、他のユーザと評価するアイテムが異なるユーザの評価データは、他のユーザとの関わりが弱く、相対的な嗜好を抽出するのには向かない。相対的な嗜好を抽出するのに向くユーザは、他のユーザも評価するアイテムを数多く評価するユーザという仮説に基づいてユーザの選択手法を提案する。図2にユーザとアイテムの関係を示し、次にユーザを定量的に評価する式を示す。

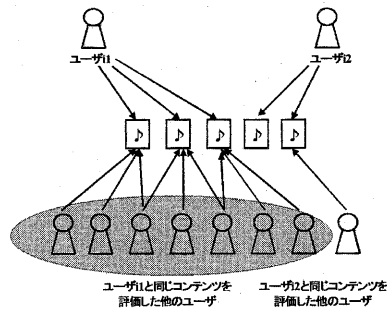


図2  
ユーザとアイテムの関係

$$u_i = \sum_{j \in m_i} \log_2 n_j \quad (3)$$

- $u_i$  : ユーザ  $i$  の重み
- $m_i$  : ユーザ  $i$  の評価したアイテムの集合

個人の評価するアイテムの数より、アイテムを評価するユーザの数が圧倒的に多いので、ユーザの

数を緩和する為にここではその対数をとる。その結果、この式によって算出されたユーザの重みによって、嗜好の抽出に必要なユーザを取捨選択をすることができる。この式によって図2におけるユーザ  $i1$  は6、ユーザ  $i2$  は1の重みをとる。

### 3.3. ユーザの潜在嗜好モデルによる嗜好の抽出

ここでは、ユーザのアイテムに対する評価の背後に隠れている嗜好を抽出する。ユーザの嗜好を抽出する目的は、協調フィルタリングの予測精度を向上することである。その為にユーザの嗜好を抽出する必要がある。ここで抽出するユーザの嗜好は、先に述べた「ヒーリング系の音楽」という嗜好に相当する因子である。我々はこの因子を抽出するためのユーザモデルをユーザの潜在嗜好モデルと呼ぶ。このモデルは、ユーザのアイテムに対する評価は、ユーザが属する因子とその因子でのアイテムの重みにより決定するという仮説に基づいている。このモデルを解くことによって、ユーザの嗜好の因子とユーザ及びアイテムの関係を数値化し、各ユーザが持つ嗜好と嗜好別のアイテムに対する重みを知ることができる。ここでの目的は因子が具体的にどのような嗜好に相当するかを知ることではなく、因子とユーザの関係を数値化することである。

次にアイテムとユーザの嗜好の関係を次のように定義する。

$h_{kj}$ : 嗜好  $k$  の因子におけるアイテム  $j$  の重み

ユーザの嗜好とユーザの関係を次のように定義する。

$p_{ik}$ : ユーザ  $i$  が嗜好  $k$  の因子に属する確率

ここで  $p_{ik}$  は  $\sum_{k \in \Omega} p_{ik} \leq 1$  を満たす。  $\Omega$  は因子の集合である。

これはユーザ  $i$  が複数の嗜好の因子に確率的に属することを意味する。

3.1 で定義したアイテム毎の重みは、  $w_k$  と  $p_{ik}$  を結合することで次のように表すことができる。全ユーザ数を  $n$ 、全アイテム数を  $m$  とする。

$$w_{ij} = \sum_{k \in \Omega} p_{ik} h_{kj} + p_{ie} h_{ej} \quad (4)$$

ここで  $p_{ie}$  はユーザ  $i$  が独自、つまりどの嗜好にも相当しない  $e$  の因子に属する確率で、  $h_{ej}$  は独自因子  $e$  におけるアイテム  $j$  の重みである。  $p_{ie}$  と  $h_{ej}$

を導入した理由は、  $p_{ik}$  と  $h_{kj}$  の結合だけで  $w_{ij}$  を説明できない誤差のためである。この誤差がユーザ  $i$  のアイテム  $j$  に対する重み  $w_{ij}$  において冗長かつ雑音になる。観測された変数に潜在変数を導入したユーザモデルに Two way aspect model [10] がある。提案手法との相違はこの独自因子の項の有無にある。提案方式は独自因子の存在により、モデルの自由度が増し、汎用性に優れ、ユーザの嗜好を抽出することができるようになると考えられる。この式をユーザ  $i$  について整理した次に示す。

$$x_i = \Lambda z_i + \varepsilon_i \quad (5)$$

$x_i$ : ユーザ  $i$  のアイテム  $j$  に対する重み  $w_{ij}$  を要素とする  $m \times 1$  の列ベクトル

$\Lambda$ : 嗜好  $k$  の因子におけるアイテム  $j$  の重み  $h_{kj}$  を要素とする  $m \times o$  の行列

$z_i$ : ユーザ  $i$  の嗜好  $k$  の因子に属する確率  $p_{ik}$  を要素とする  $o \times 1$  の列ベクトル

$\varepsilon_i$ : ユーザ  $i$  の独自因子の積  $p_{ie} \times w_{ej}$  を要素とする  $m \times 1$  の列ベクトル

更にこの式を全ユーザについて整理した式を次に示す。

$$X = Z\Lambda^T + \varepsilon \quad (6)$$

$X$ : ユーザ  $i$  のアイテム  $j$  に対する重み  $w_{ij}$  を要素とする  $n \times m$  の行列

$Z$ : ユーザ  $i$  の嗜好  $k$  の因子に属する確率  $p_{ik}$  を要素とする  $n \times o$  の行列

$\Lambda^T$ :  $\Lambda$  の転置行列

$\varepsilon$ : ユーザ  $i$  の独自因子の積  $p_{ie} \times h_{ej}$  を要素とする  $n \times m$  の行列

全ユーザについてのユーザモデルはこのような行列表現で整理できる。この行列は、  $p_{ik}$  を制約条件として EM アルゴリズム[4]を用いて解析的[5]に解くことができる。

### 3.4. ユーザの評価の予測手法

ここではユーザのアイテムに対する評価の予測を行う式を提案する。ユーザ  $i$  のアイテム  $j$  に対する評価を予測する場合、アイテム  $j$  を評価した全ユーザについて次の行列を求める。

$$X_j = Z\Lambda_j^T + \varepsilon_j \quad (7)$$

$X_j$ : アイテム  $j$  を評価したユーザのアイテムに対する重み  $w_{ij}$  を要素とする  $n_i \times m$  の行列  
 $\Lambda_j^T$ : 嗜好  $k$  の因子におけるアイテム  $j$  の重み  $h_{kj}$  を要素とする  $m \times o$  の行列の転置行列  
 $\varepsilon_j$ : アイテム  $j$  を評価したユーザ  $i$  の独自因子の積  $p_{ie} \times h_{ej}$  を要素とする  $n_i \times m$  の行列  
 $X_j, \varepsilon_j$  は評価データからアイテム  $j$  を評価したユーザについて抽出し、 $Z$  は先に式(6)から求めたものをそのまま用いる。よって(7)より  $\Lambda_j^T$  は一意的に算出することができる。次にこの新たに算出した  $\Lambda_j^T$  の要素  $h_{kj}$  を用いて、ユーザ  $i$  のアイテム  $j$  に対する評価の予測  $\hat{r}_{ij}$  を求める式を次のように定義する。

$$\hat{r}_{ij} = \frac{1}{g_j} \sum_{k \in \text{keo}} p_{ik} h_{kj} + p_{ie} h_{ej} \quad (8)$$

## 4. 評価実験

### 4.1. 評価データ

ここでは提案手法の有効性を予測精度とユーザ選択の比較実験により確認する。実験に用いるユーザの評価データは Each Movie Data Set である。これは Compaq Equipment Corporation[12] によって運営されているサイトより作成され、72916 ユーザによって 1628 の映画に対しての評価から構成されている。評価方法はユーザが映画に対して 0-1 の間、間隔 0.2 の数値を与えて評価している。

このデータを利用した Breese[1] や Canny[3] の実験は、無作為に抽出した 5000 人のユーザの評価を用いた。我々の実験では、テストデータは同数のユーザを利用するが、学習データはユーザの数を 72916 人から選択し、順次減少させて用いる。

ここで Each Movie Data Set の基礎統計を示す。図 4.1 は映画毎に評価したユーザ数と映画の数の関係である。横軸は映画別の評価したユーザ数で、縦軸は評価したユーザ数別の映画の数である。図 4.1 は全体の 71.1% の映画は 100 人以上のユーザに評価され、ユーザ間の関わりがあることを示している。

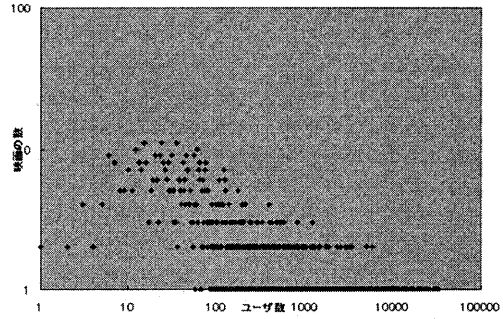


図 4.1 映画を評価したユーザの数と映画の数

図 4.2 はユーザ毎に評価した映画の数とユーザの数の関係である。横軸はユーザの評価した映画別の数で、縦軸は評価した映画数別の映画の数である。図 4.2 は全体の 74.1% のユーザは 1~100 本の映画を評価していることを示している。図 4.1 と比べて、ユーザの評価した映画の数は映画別のユーザ数よりその分布は小さいことが確認できる。従って、ユーザを取捨選択する時、映画を評価したユーザの数をそのまま使うと、その影響が大きくなる。そこで提案手法のように緩和することで嗜好の抽出に適切なユーザの重み付けができると考えられる。

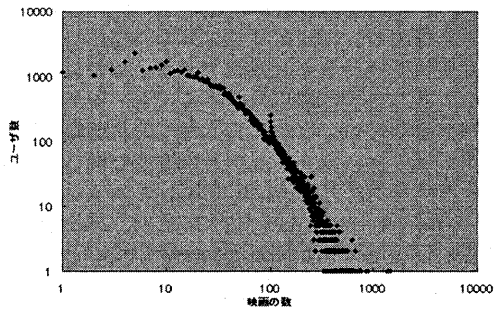


図 4.2 ユーザが評価した映画の数とユーザの数

### 4.2. 評価方法

評価方法として Breese らの Allbut1[1] を用いた。Allbut1 はユーザのアイテムに対する評価の 1 つをマスクして、テストデータとする。それ以外のユーザの評価を学習データとして、マスクしたアイテムに対するユーザの評価を予測する。我々はこのマスクをユーザ毎で最新の評価に対して行った。

### 4.3. ユーザを選択する方法

ユーザを選択する方法は、無作為、ユーザの評価した映画の数、提案したユーザの重み付けを用いる三通りの方法である。無作為以外の方法はそれぞれの値によってユーザを順位付けし、その順位が低いユーザを取り除くことでユーザの選択を行った。

### 4.4. 精度を比較する方法

予測精度を比較する手法として、映画別の平均点、ピアソンの積率相関 (significance weighting method を併用) を用いた方法[6]を用いた。映画別の平均点は、ユーザ別の映画別の評価を予測するのではなく、各映画の平均点を予測として用いるものである。予測精度を評価する基準として本実験では Mean Absolute Error (MAE) を利用する。

### 4.5. 実験結果とその考察

以下、評価実験の結果を示す。図 4.3 はユーザ数を無作為抽出、ユーザ毎の映画の数、提案手法の三手法で減少させた場合のピアソンを用いた MAE の結果で、図 4.4 は手法毎の MAE の比較、表 4.1 は任意に抽出した二つの因子において、重み  $h_{kj}$  が高い映画を上位五件示したものである。図 4.3 より、ユーザ選択手法を比較すると、評価した映画の数と提案手法の場合、ユーザの数が 30000 までは大きな差異は認められないが、それ以降は評価した提案手法以外の MAE の増加率が上昇していることが確認できる。これは計算に用いたユーザの映画の数は多くても、他のユーザとの重なりが小さい為に、予測の精度が落ちた為と考えられる。図 4.4 より予測精度に関しては、提案手法の方が低い MAE を示していることが確認できる。表 4.1 では各因子において重みの高い映画より、その因子の嗜好を考察することが出来る。この結果より、提案した潜在嗜好モデルはユーザの嗜好を抽出することが出来ると考えられる。

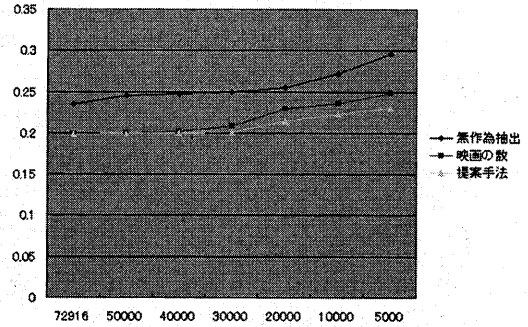


図 4.3 ユーザ選択による MAE の変化

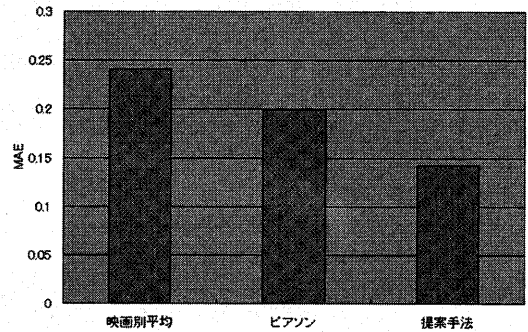


図 4.4 提案手法別の MAE の結果

表 4.1 任意に抽出した因子において重みの高い映画上位五件

因子 k	因子 k'
Jumanji	Grumpier Old Men
How to Make an American Quilt	Casino
Toy Story	Assassins
Powder	The American President
Sabrina	Tom and Huck

以上の実験結果より、提案手法は、ユーザの嗜好を抽出することで精度の高い予測ができ、計算量の低減に有効なユーザの選択手法であることが確認できた。

## 5. まとめ

本研究は、協調フィルタリングにおいて予測精度が高く計算量を低減した新しい手法を提案した。提案手法を Each Movie Data Set に適用した結果、従来よりもユーザの評価を少ない計算量で高い精度で予測ができることを確認した。今後はユーザの選択と予測精度の定量的な関係を明らかにすることを課題とする。

Mouth'," Proceedings of the CHI-95 Conference, Denver, CO, ACM Press, May 1995.

[12] Each Movie:

<http://research.compaq.com/SRC/eachmovie/>

## 文 献

- [1] J. S. Breese, D. Heckerman and C. Kadie: Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pp 43-52, July 1998.
- [2] J. Canny: Collaborative Filtering with Privacy, IEEE Conf. on Security and Privacy, Oakland CA, May 2002.
- [3] J. Canny: Collaborative Filtering with Privacy via Factor Analysis, Proceedings of the 25'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere Finland, August 2002.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm Journal of the Royal Statistical Society, B, 39, 1-38. 1977.
- [5] Z. Ghahramani and G.E. Hinton: The EM Algorithm for Mixtures of Factor Analyzers University of Toronto Technical Report CRG-TR-96-1.
- [6] J. L. Herlocker, Joseph A. Konstan, Al Borchers, John Riedl: An algorithmic framework for performing collaborative filtering, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1999), pp.230-237, Berkeley, California, United States, August 1999.
- [7] W. Hill, L. Stead, M. Rosenstein and G. Furnas: Recommending and Evaluating Choices in a Virtual Community of Use Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'95), pp.194-201, May 1995.
- [8] D. Pennock and E. Horvitz, S. Lawrence, and C. L. Giles: Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach, Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000), pp.473-480, Stanford, CA, June 2000.
- [9] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl: GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In CSCW '94: Conference on Computer Supported Cooperative Work, ACM, pp. 175-186. Chapel Hill, NC, Oct 22-26, 1994.
- [10] A. I. Schein, A. Popescul, L. H. Ungar and D. M. Pennock: Methods and Metrics for Cold-Start Recommendations. Proceedings of the 25'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, Aug 2002.
- [11] U. Shardanand and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of