

## 目利き利用者の発見と利用によるソーシャルフィルタリング方式

3R-4

有吉 勇介 市山 俊治

NEC 関西 C&amp;C 研究所

E-mail:{ariyoshi, ichiyama}@obp.cl.nec.co.jp

## 1.はじめに

近年の情報の氾濫はひどく、最近の WWW 等の流行はさらにそれに拍車をかけている。情報フィルタリング技術とはこのような情報の洪水の中から利用者の興味や嗜好に合う情報を見つけ出すための技術である。従来からキーワードや単語出現頻度等を使った全文検索を行う Content Based Filtering(以下 CBF)方式が研究されてきた。しかし、CBF は単語の出現情報に特徴があらわれる技術文書等に向いており、それ以外の情報に対してはフィルタリングの精度が下がるという欠点があった。

しかし最近、このような欠点の無い Collaborative Filtering もしくは Social Information Filtering(以下 SIF)と呼ばれる手法が研究されている。例えば[1]では NetNews を対象にした SIF システム GroupLens について、[2]では音楽アルバムを対象とした SIF システム Ringo について報告している。

SIF は興味や嗜好の似た別の利用者が良いと評価した情報を推薦する方式である。SIF では利用者はシステムから情報の推薦を受けると、その情報の趣味・嗜好に合う・合わないを評価しシステムに返す。一方システムが利用者に情報の推薦をする時は、そのフィードバック評価パターンの似た別の利用者を探し、似た利用者が良い評価をしている情報を推薦する。

このように SIF は利用者の情報に対する評価だけを利用するため、次のような特徴がある。

1. 利用者は評価フィードバックするだけで良く、検索式やプロフィールを書かなくて良い。

2. 技術文書以外の情報、文芸的文書や文書以外のメディアに適用できる。
3. 評価数量が増えるほどフィルタリングの精度が向上する。

しかし SIF には次の2つの欠点がある。1つ目の欠点は新しい情報に対するフィルタリング精度が低いことである。これはフィルタリング精度が集まった評価数量に依存するという SIF の3番目の特徴に起因している。つまり、新しい情報は評価している利用者が少ないので、それに応じてフィルタリング精度が低くなってしまふ。2つ目の問題は興味・嗜好が似ている利用者が複数人いないと推薦できないことである。次節で説明する従来の SIF 方式自体は複数の利用者が評価していれば、それが要求者と似ていない利用者であっても情報の推定評価は求まる。しかし実際のシステムでは要求者と似ている利用者が複数人居ないと精度が低いので推薦を行わず、似ている利用者が複数居るときだけ推薦することが多い。

そこで、ここでは目利き利用者、つまり要求者の興味や嗜好に合った情報を誰より早く見つける利用者を発見しそれを活用することによりこれらの欠点を軽減する方式を提案する。目利き利用者が発見できれば、評価が少ない新しい情報に対しても目利き利用者の評価を使ってフィルタリング精度を高めたり、目利き利用者が評価したかどうかで要求者に推薦するかどうかを決めることが出来る。

## 2.従来の SIF 方式

従来の SIF は次のようにモデル化できる。今、情報を推薦する対象である利用者（以下、要求者と呼ぶ）に対して、システムは次の式より要求者の各情報に対する評価を推定する。システムには  $m$  個の情報と  $n$  人の利用者があり、便宜上1番目の利用者が要求者として考える。

・  $v_{ki}$ :  $k$  番目の利用者の  $i$  個目の情報に対する評価

A Social Information Filtering Method with Finding Leading-edges.

Yusuke Ariyoshi, Syunji Ichiyama

Kansai C&C Research Laboratories, NEC corp.

4-24, Shiromi 1-Chome, Chuo-ku, Osaka 540, Japan

- $P_k = \{v_{k1}, \dots, v_{km}\}$ : k 番目の利用者のプロフィール・ベクトル
- $F(P_1, P_k)$ : 要求者と k 番目の利用者の類似度
- $E_1$ : 要求者の推定評価ベクトル

$$E_1 = \frac{\sum_k F(P_1, P_k) P_k}{\sum_k F(P_1, P_k)}$$

この式は要求者との類似度で重みを付けたプロフィールの平均を求めていると解釈することもできる。普通、類似度は要求者と k 番目の利用者のプロフィールベクトルの相関係数が用いられる。[2]では相関係数の改良について論じられている。

### 3. 利用者の目利き度

ここでは各利用者の目利き具合を示す目利き度を定義する。目利き度は有効な情報をより早く見つける性質を表すように、要求者と似た評価をより早くするほど大きくなるようにすべきである。そのため評価情報に加えて、評価をいつしたかという評価時刻情報を使い、以下の式で目利き度を計算する。

- $L_k$ : 要求者からみた k 番目の利用者の目利き度
- $e_{1l}$ :  $E_1$  の 1 番目の要素、1 番目の情報に対する評価の推定値
- $f(e_{1l}, e_{kl})$ : 要求者と k 番目の利用者の 1 番目の情報に対する評価の類似度
- $N_l$ : l 番目の情報を評価している利用者の人数
- $t_{1l}$ : l 番目の情報が最初に評価された時刻
- $t_{lk}$ : k 番目の利用者が l 個目の情報を評価した時刻
- $t_{now}$ : 現在時刻

$$L_k = \frac{\sum_l 2^{-\frac{N_l (t_n - t_k)}{t_n - t_{now}}} f(e_{1l}, e_{kl})}{\sum_l 1}$$

この式では各情報についての要求者と k 番目の利用者の評価の類似度を平均しているが、類似度に k 番目の利用者の評価時刻が遅いほど指数的に小さくなる係数を掛けてから平均を取っている。それにより有効な情報をより早く見つける性質を示すものになっている。

### 4. 目利き利用者を活用した SIF

前項の目利き度を使い、従来の SIF の適合度計算式を改良したものを下の式で示す。

$$E_1 = \frac{\sum_k L_k F(P_1, P_k) P_k}{\sum_k L_k F(P_1, P_k)}$$

この式はプロフィールベクトルの平均を要求者との類似度と利用者の目利き度で重み付けしたと解釈することもできる。これにより従来より精度の高い推薦ができ、特に従来精度の低かった新しい情報に対する精度の向上に効果が大きい。

次に推薦するかどうか決められないという問題に対処するため、目利き利用者の評価を直接利用する。つまり、目利き度の高い利用者が新しい情報に高い評価を付けた場合、統計処理をせずに直接要求者に推薦する。この場合、評価をつけた利用者が目利き度の高い利用者ただ一人でも要求者に情報が推薦されるので、上記の問題が従来より改善される。

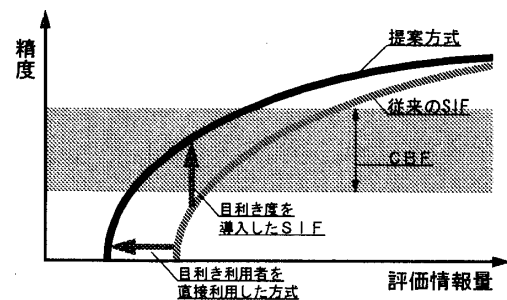


図 1: フィルタリング精度の模式図

### 5. おわりに

現在、社内で SIF の実験を行なう準備をしている。この実験で提案方式の評価を行い、それをもとにさらに改良する予定である。

### 6. 参考文献

- [1] "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", Paul Resnick, et al., Proc. CSCW'94.
- [2] "Social Information Filtering: Algorithms for Automating "Word of Mouth"", Upendra Shardanand, Pattice Maes, Proc. CHI'95.