

# セレンディピティに基づく大域的協調フィルタリング

大槻有紗<sup>†</sup> 三浦 孝夫<sup>†</sup>  
<sup>†</sup> 法政大学理工学部創生科学科

## 1. 協調フィルタリング

$N(j,a)$  のユーザ  $k$  の評価値  $r(k,j)$  を類似度  $S(x,k)$  で加重平均した値を、推定値  $r(a,j)$  すると次の(1)の式で定義される。

$$r_{aj} = b_{xi} + \frac{\sum_{j \in N(j;a)} S_{xk} \cdot (r_{kj} - b_{xi})}{\sum_{j \in N(i;a)} S_{xk}} \dots\dots\dots(1)$$

$b(x,i)$  は  $R$  の規準を仮定しており、 $b(x,i) = \mu$  (平均判定値) +  $b_x$  (商品  $x$  の平均評価値) -  $\mu$  +  $b_i$  (ユーザ  $i$  の評価値分散) で求められる。本研究による大域的であるとは、推定順序によらず推定値が同一であることを保証する値であると定義し、式(2)で表す。

$$S\vec{x} = \lambda\vec{x} \dots\dots\dots(2)$$

$S$  はユーザ  $i$  とユーザ  $j$  の正方対称行列  $S(i,j)$  を表しており、 $\lambda$  は固有値を表している。左固有値 1 とすると式 2 は  $Sx=x$  となる。 $x$  の固有ベクトルを求めることによって、協調フィルタリングの極限解を求める。極限解とは、協調フィルタリングを極限までおこなうことで得られる、極限値のことで推定順序によらず推定値が安定な解である。ゆえに極限界を出すことにより、大域的な値を検出することができる。

## 2. 提案手法

本研究では、セレンディピティな推薦手法を提案する。セレンディピティとはユーザにとって物珍しく興味のあるものでなければならない。セレンディピティは新規性・関連性・意外性の三つの手法を組み合わせることにより推薦できると考える。

新規性はユーザが一度も見たことや聞いたことがないものと定義する。 $n$ 名のユーザによる  $m$ 種類の項目評価を行列  $D(n \times m)$  で与えると各項目値  $D(u,i)=v$  はユーザ  $u$  が項目  $m$  に評価値  $v$  を与えたことを表している。5段階評価の場合  $v=1, \dots, 5$  となり、未定義なら 0 とする。 $D(u,i)$  が未評価、すなわち  $D(u,i)=0$  ならば、ここでは協調フィルタリングで推定した値  $v_0$  をそのまま利用する。未知の度合いを示すため、次の(3)の様で定義する。

$$\bullet \text{ novel}(u, i) = \frac{|v_0 - v_1|}{5} \dots\dots\dots(3)$$

関連性はユーザ  $u$  が項目  $i$  に有する興味と定義する。新規性と同様に、評価値  $D(u,i)$  が未定義のとき、協調フィルタリングによる推定値  $v_0$  を仮定する。 $D(u,i)=v_1$  が与えられているとき、 $v_0$  との差  $|v_1 - v_0|$  は推定値  $v_0$  との評価差をあらわし、(4)で与える。

$$\bullet \text{ relevace}(u, i) = 5 - |v_1 - v_0| \dots\dots\dots(4)$$

意外性は  $D(u,i)$  または協調フィルタリングによる推定値が、他ユーザの評価とどの程度ばらついているかを評価し、この値を  $unexp(u,i)$  とする。ユーザ  $u$  の項目  $i$  の評価値  $D(u,i)$  が、ユーザ  $j=1, \dots, n$  との差異を  $|D(u,i) - D(j,i)|$  の 2 乗であらわす。このときこれを類似度  $sim(u,j)$  による重み付き和を  $unexp(u,j)$  とすると意外性は(5)のように定義される。

$$unexp(u, i) = \frac{T(u, a)}{S(u)} \dots\dots\dots(5)$$

(5) は、ユーザ  $u$  との類似度合計  $S(u)$ 、類似度による重み付きした差異の和  $T(u,a)$  を表している。

セレンディピティ  $serendipity(a,b,c)$  は(3)(4)(5)の調和平均であ

り、(7)で定義される。

$$serendipity = \frac{3abc}{ab + bc + ac} \dots\dots\dots(6)$$

(6) は新規性、関連性、意外性を同時に評価し1つにまとめた値であるが、同じだけの重要性を表している。

## 3. 実験

実験データは、NII 楽天トラベルデータ全 341 件から、目的変数評価 1(立地)評価 2(部屋)評価 3(食事)評価 4(風呂)評価 5(サービス)評価 6(施設・アメニティ)評価 7(総合評価)価格を用いる。

本実験のセレンディピティであるかの評価は、まず各ユーザごとにデータを比較し行う。その後各アンケートデータのロコミ内容を点数化し結果と比較し正当性を確認する。苦情のデータ場合 no・文に良かった内容が書いてある場合加点+1・文に不満・改善点が書いてある場合減点-1として、点数化する。

セレンディピティなデータはユーザ 233 とユーザ 49 の 2 件になる。



図1 コメント点数化

ユーザコメントから見てもユーザ 1 とは全く違うホテル、そしてユーザ 1 より評価平均点数が約 1~2 点高くなっており、総合的に良いホテルを推薦している。しかし、ユーザ 1 と関連性あり(家族向け)、点数がただ高いユーザを推薦していない(最大値は 8 点だが、推薦値は 5 点・4 点)という特徴がある。

セレンディピティを新規性・関連性・意外性を組み合わせてセレンディピティを測定すると、ユーザ 1 とは全く違うホテル、そしてより総合的に良いホテル(評価点数の平均が約 1~2 点高い)を推薦している。しかし、ユーザ 1 と関連性あり(家族向け)、点数がただ高いユーザを推薦していない。ゆえに、ユーザ 1 のセレンディピティな推薦ができていると考える。よって他のユーザのデータもセレンディピティな推薦ができている。

## 4. 結論

本研究により、セレンディピティは意外性・新規性・関連性の 3 つの指針の組み合わせることにより表せることがわかった。そして、セレンディピティなデータの推薦することができた。

参考文献 [1] Denis Kotkov, Shuaiqiang Wang, Jari Veijalainen (2016)「A Survey of Serendipity in Recommender Systems」