情 報 学 基 礎 51－3
自 然 言 語 処 理 127－3
（1998. 9. 17）

多変量解析を用いたソーシャル情報フィルタリング

有吉 勇介　市山 俊治
NEC ヒューマンメディア研究所
E-mail:{ariyoshi, ichiyama}@hml.cl.nec.co.jp

他者の評価を利用した情報フィルタリング方式であるソーシャル情報フィルタリングでは、従来は利用者の関心の類似度として評価履歴間の相関係数を用い、利用者の情報に対する評価値を、他者の評価を利用者間類似度を重みとする重み付平均で予測していた。本発表では相関係数の代わりに多変量解析手法の一つである双対尺度法を使った方式を提案し、実験結果を紹介する。従来の相関係数は利用者の関心がどれだけ違うかという距離情報であったが、本方式はどう違うかという方向情報も利用するためフィルタリング精度が高い。また、従来は利用者間類似度が閾値以下のものは切り捨てていたため予測評価値が算出できない情報があったが、本方式は切り捨てを行なわないため全情報について評価値予測が出来る。

Social information filtering using multi-variate analysis

Yusuke Ariyoshi, Syunji Ichiyama
Human Media Research Laboratories, NEC corp.
E-mail:{ariyoshi, ichiyama}@hml.cl.nec.co.jp

Social information filtering (SIF) is a filtering method which uses score of information by other user. Conventional SIF method uses correlation coefficient for similarity between users, and weighted average for score estimation. In this report we propose a new SIF method using dual scaling which is a method of multi-variate analysis In new method, similarity and score estimation are multi-dimension, and does not cut off score estimation by threshold of similarity. we test and compare two methods. As a result, new method is lower error rate than conventional method.

1. はじめに

インターネットの普及により、個人が世界中の情報にアクセスし、また世界に向けて情報発信することが簡単にできるようになった。しかしそのため情報の氾濫はいっそうひどくなり、自分に必要な情報の選別がますます大変な作業になっている。このような情報洪水の中から利用者が必要とする情報を選別する技術が情報フィルタリングである[1,4]。

情報フィルタリングは、利用者の情報に対する評価を予測することで情報の選別を行う。評価予測方式には、キーワードや単語頻度を使う内容に基づく(Content-based)方式と、情報に対する他の利用者の評価を利用するソーシャル情報フィルタリング（SIF）方式[1]の２つに大きく分けられる。

SIF はキーワード等の情報の中身を使用せず、利用者からの評価だけを使って評価予測するため、内容に基づく方式では苦手だった文芸的文書や画像・音楽などのマルチメディアコンテンツの選別も可能である。しかし、従来の SIF アルゴリズムは予測精度があまり高くない。そのため利用者を関心が似ている者だけに絞りこんで精度を高めるが、そうすると評価予測できない情報が増えてしまうという課題があった。

従来の SIF アルゴリズムは利用者間類似度として評価履歴の相関係数を用い、相関係数を重みとして他者の評価の重み付平均によって評価予測を行っていた。ここでは利用者間類似度として従来方式で使われていた評価履歴の相関係数の代わりに、多変量解析手法を使用する方式を提案し、実験結果を紹介する。従来方式は利用者間で関心がどれだけ違うかという距離だけを使っていたが、提案方式では多次元空間を使用するため関心がどう違うかという方向情報も利用できるので予測精度が高くなっている。また、利用者の絞り込みを行わないため、全情報について評価予測できる。

[1] 協調(Collaborative)フィルタリングとも呼ばれる

2. SIF:ソーシャル情報フィルタリング

ＳＩＦのコンセプトは、人間が日常行なっている面白い情報を同じ興味を持つ仲間同士で教え合うというロコミによる情報伝達のシステム化である。

ここでは、利用者が情報に対して数値による評価を返す SIF を対象にしているが[2]、このタイプの SIF はインデックス生成と評価予測の２段階で構成される[2]。

インデックス生成：

従来の SIF 方式では、各利用者はインデックスとして他の利用者との興味の類似度（利用者間類似度）のリストを持っている。この利用者間類似度としては、２人の評価履歴のうち共に評価している情報に関する部分を抜き出し、その相関係数を計算して使っている。

評価予測：

情報ａに対する評価の予測は、情報ａを評価済み利用者が付けた評価値を、利用者間類似度を重みとした重み付平均によって計算する。

MIT の Ringo[3]では相関係数の計算で使用する平均値を、計算せずに評価の真ん中の値（５段階評価ならば３）で代用している。また相関係数が閾値未満の利用者は予測に用いないようにして予測精度を高めている。

3. 従来方式の課題

SIF はロコミをシステム化したもので、人により内容がチェックされた情報だけが推薦されるので、単にキーワード等で検索した情報より、推薦の精度が高いと言われている。しかし、従来の SIF 方式には次の課題がある。

- 期待されていたほど予測精度は高くない。内容に基づいた方式とほぼ同じという実験結果もある[5]。

[2] 評価を文章で返すものや、情報を知らせてあげたい利用者のアドレスを返す SIF もある。

- 他の利用者が評価済みの情報であっても一部の情報は評価予測できない

従来の SIF 方式の精度が低い原因として次のような理由が考えられる。

1）相関係数計算に評価値をそのまま使用

評価値には、たとえば1から5までの5段階の段階評価値を使用しているが、本来、段階評価値は数値の順序のみが意味を持つ。例えば3つの情報（ア・イ・ウ）にそれぞれ1、2、3と評価値を付けた場合、アよりイ、イよりウの評価が高いということを表わす。しかし数値の差に意味はなく、アとイの差はわずかであるがイとウの差は非常に大きいことが有り得る。ところが、従来の評価値をそのまま使って相関係数を計算するということは、評価値の差にも意味があり、各評価段階は等間隔であることを仮定しているが、この仮定には矛盾がある。

2）利用者間類似度に情報の直交性を仮定した相関係数を使用

相関係数は、幾何的には個々の情報の特徴が直交していることを仮定した場合の、評価履歴ベクトルがなす角のコサイン値である。しかし、情報同士は似ているものもあれば似ていないものもあり、直交性の仮定は成り立たない。

3）インデックス（利用者間類似度）が1次元

例えば音楽を対象としたとき、ある利用者に対して利用者AとBは利用者間類似度が同じであっても、Aはロック好きだがはクラシック好き、といった興味の方向性があるはずである。しかし、従来の SIF 方式はインデックスとして利用者間類似度という一次元の距離情報を使用しているため、興味の方向性を評価予測に利用していない。

4）予測方式（重み付平均）が単純

相関係数を重みとして、評価の重み付き平均により評価予測を行なっている。これは幾何的には直線当てはめにより評価予測を行なっている。しかし、2次式等のより高度な予測式を用いれば精度が向上することが期待できる。

5）相関係数が閾値以下の利用者の評価情報が未使用

従来方式は相関係数が閾値以上の利用者の評価情報だけを利用しているため、閾値未満の利用者がした評価は利用されない。

また、一部の情報は評価予測できない理由も5)と同じである。従来方式は相関係数が閾値以上の利用者の評価情報だけを利用しているため、相関係数が閾値未満の利用者しか評価していない情報は評価予測出来ない。

4. 提案方式

従来の SIF 方式の課題に対する上記の考察に基づき、提案方式を以下の指針に従って設計した。

4.1. 方針

♦ 段階尺度から順位尺度へ変換

段階評価値を、各利用者の評価値の付け方に合わせて、差に意味がある数値に変換してからフィルタリングを行う。

♦ インデックス：情報空間の導入

提案方式では次の性質を持つ情報空間を多変量解析により生成してインデックスとして使用する。

・情報空間上での情報と利用者の配置方針
- 情報間について、利用者による評価が似ている情報同士ほど近くなるように配置
- 利用者間について、情報に対する評価が似ている利用者同士ほど近くなるように配置
- 利用者－情報間について、利用者に高く評価された情報ほどその利用者の近くに配置

このように情報空間は情報の直交性という非現実な仮定はしてない。また、情報空間をインデックスに用いることにより、評価予測に距離だけでなく方向も利用できるようになる。

♦　評価予測：２次曲面による予測

提案方式では情報空間上で２次曲面によって評価予測をする。そのために、まず情報空間上での利用者と情報と評価の関係を分析して２次曲面による関心モデルを算出する。この関心モデルを使って未評価の情報の予測評価値を計算する。

このように多次元空間上で２次曲面を使って評価予測することにより、従来の重み付平均より高度な予測が行なえ、１次元の距離情報だけでなく興味の方向性も利用できるようになる。また、利用者間類似度に閾値を設けて閾値未満の利用者の評価情報が利用されないというようなことは行わないため、従来は評価予測が出来なかった情報に対しても評価予測できるようになる。

4.2.　多変量解析を用いたSIF

ここでは上記の方針に基づいて設計した提案方式の説明を行なう。提案方式は、尺度変換、情報空間生成、関心モデル生成、評価予測の４段階で構成される。中心となる情報空間生成は多変量解析の双対尺度法[5]によって行なう。また尺度変換も[5]に記載されている方法を改良したものである。

4.2.1.　段階尺度から順位尺度へ変換

段階評価を値の差に意味のある順位尺度による数量に変換するために、境界点の導入と順位付け、同点の処理の３つを行なう。

以下では例として表１の５人の利用者が３件の情報に対して３段階評価したものを使う。ただし、表１では利用者が評価しなかった場合を＊で示している。

表 1：評価例

利用者＼情報	C1	C2	C3
U1	＊	3	2
U2	2	3	2
U3	3	＊	3
U4	＊	3	2
U5	2	2	＊

・境界点の導入

ここで、各評価段階の境界点Ｓを導入する。例は３段階評価なので境界点はＳ１、Ｓ２ の２つになる（図１）。

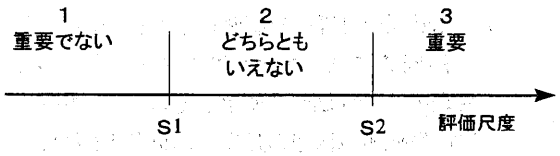

図 1：評価値と境界点

・順位付け

次に、情報と境界点を合わせて順位付けする。このとき利用者が異なる情報に同じ評価値を付けた場合は可能性のある順位を全て並べる。表２は表１を各利用者ごとに順位付けしたものである。

表 2：評価例（順位）

利用者＼情報	C1	C2	C3	S1	S2
U1	＊	1	3	4	2
U2	3,4	1	3,4	5	2
U3	1,2	＊	1,2	4	3
U4	＊	1	3	4	2
U5	2,3	2,3	＊	4	1

・同点の処理

異なる情報に同じ評価が付いている場合、順位付けで書き並べた順位を平均した順位で代表させる。表３は、表２に同点の処理を行なっ

表 3：評価例（平均化順位）

利用者＼情報	C1	C2	C3	S1	S2
U1	＊	1	3	4	2
U2	3．5	1	3．5	5	2
U3	1．5	＊	1．5	4	3
U4	＊	1	3	4	2
U5	2．5	2．5	＊	4	1

表 4：中心化

利用者＼情報	C1	C2	C3	S1	S2
U1	*	3	−1	−3	1
U2	−1	4	−1	−2	2
U3	2	*	2	−5	−1
U4	*	3	−1	−3	1
U5	0	0	*	−3	3

たものである。

このようにして段階評価を順位尺度に変換することで、数量の差は順位の差という意味を持つ。また順位尺度化すると「この利用者は5をほとんど付けない評価の厳しい利用者た。」といった段階評価が持っている情報が失われてしまうが、境界点を導入することでそれを防いでいる。

4.2.2. 情報空間生成（双対尺度法）

情報空間を生成するために、数量の中心化、比較判断数の導入、双対尺度法による情報空間生成の3つを行なう。

・中心化

双対尺度法を行なう前処理として、順位尺度に変換した評価データを各利用者の順位平均が0になるように中心化を行なう。表4は表3を中心化したもので、具体的には表3から利用者毎に数量を（順位中心−順位）＊2している。

・比較判断数の導入

比較判断数とは、利用者がある情報の順位を決めるときに、何個の別の情報と比較して決めたかを表す。例えば、利用者が4つの情報の順位を決めた場合、ある情報の順位は残りの3つとの比較判断に基づいてきまる。つまり、比較判断数は評価済み情報数−1となる。表5は例の比較判断数であり、右と下の端に和も記してある。

表 5：比較判断数

利用者＼情報	C1	C2	C3	S1	S2	比較判断数 fn=Dn
U1	*	3	3	3	3	12
U2	4	4	4	4	4	20
U3	3	*	3	3	3	12
U4	*	3	3	3	3	12
U5	3	3	*	3	3	12
f=D	10	13	13	16	16	ft=68

・双対尺度法

以上のデータから双対尺度法を用いて情報空間を生成する。行列Fを表4の値を要素に持つ行列をFとする。ただし未評価を示す＊は0に代えてある。行列Dは対角要素に比較判断数 f を持つ対角行列、0行列Dn は対角要素に比較判断数 fn を持つ対角行列とする。また、行列Cは以下の式で定義される行列とする。

$$C = D^{-1/2} F^t D_n^{-1} F D^{-1/2}$$

すると、双対尺度法による情報空間の生成は以下の固有値η固有ベクトル w の固有方程式を解く問題に帰着される。

$$(C - \eta^2 I)w = 0$$

ことのき、各情報の情報空間上での座標を表すベクトルxは以下の式で求められる。

$$x = D^{-1/2} w$$

また、各利用者の座標を表すベクトルyは以下の式で求められる

$$y = \frac{D_n^{-1} F x}{\eta}$$

4.2.3. 関心モデル生成

情報空間上での利用者と情報と評価の関係を分析して関心モデルを生成する。ある利用者の評価値を情報空間で情報の上に表示した散布図が図2である。数値が評価値、点が未評価情報、右端の大きな点が評価を付けた利用者である。関心モデル分析では、直観的には散布図上での利用者と情報とその評価値の関係をあらわす式を求めることが目的である。

・回転・重ね合わせ

利用者一人の評価データだけではデータ数が少なく関心モデルの精度が低くなるため、各利用者の散布図を重ね合わせてデータ数を増やす。重ね合わせる際は、原点から利用者への方向が同じになるように散布図（情報空間）を回転させてから重ね合わせる。また、図２では分かりやすくするために評価値をそのままプロットしているが、実際の処理では表４の中心化順位を用いる。

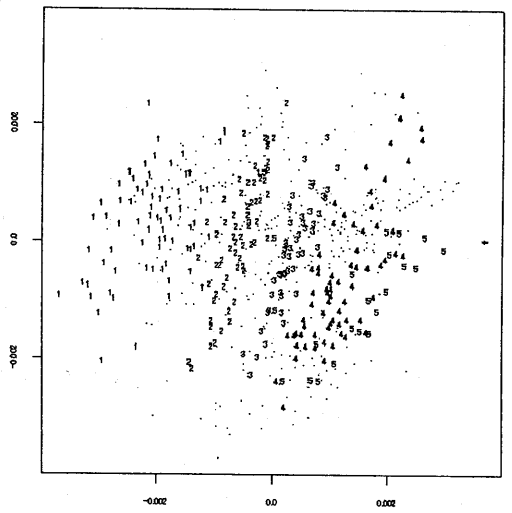

図 2：情報空間上の評価値

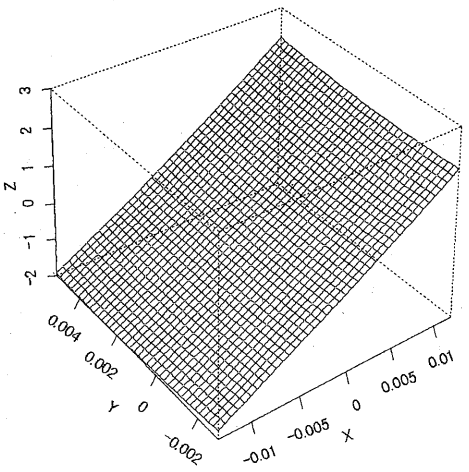

図 3：関心モデル

・関心モデル生成

回転・重ね合わせしたデータから、２次曲面を表わす式の各項の係数を最小二乗法で求める。図３は関心モデルの２次曲面の鳥瞰図で、ＸＹ平面が情報空間でＺ軸が中心化順位を示している（Ｚ軸は±３に入るように正規化してある)。

4.2.4. 評価予測

未評価の情報の評価値予測を行なうには、まず関心モデル生成での回転と同様に、情報空間の原点から注目している利用者への方向と関心モデルの原点から利用者への方向を回転によって合わせる。つぎに、未読情報の座標値を関心モデルに代入して中心化順位の予測数量を求める。予測数量を、段階評価へ逆変換することで予測評価値が求められる。

5. 評価実験

5.1. 実験データ

実験では９０人の利用者に技術文書のフィルタリングを行ない評価データを収集した。利用者には技術文書を読んで関心に応じて１(なし)から５(あり)の５段階評価をしてもらった。約１２０００件の評価を受け取った。評価実験

表 6：評価分布

評価数	評価点数分布				
	1	2	3	4	5
4651	1703	816	750	738	468

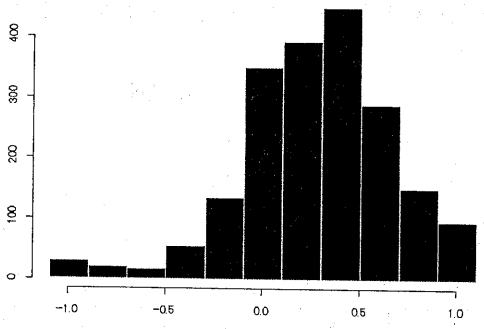

図 4：利用者間類似度

ではそのうち文書を８人以上に評価されている３６１件にしぼり、利用者は１７件以上評価した４５人に絞った。評価データは４６５１件で、利用者当たりの平均評価数は１０３．４件／人、文書あたりは平均１２．８８件／文書となり、評価データの行列中で分かっているエントリの割合は２７．６％である。評価点の分布を表６に示す。利用者間類似度のヒストグラムは図４の通りである。

5.2. 実験方法

実験は評価データを１０ブロックに分割し、１つのブロックを残り９ブロックのデータから予測することを、全１０ブロックについて行ない、平均を取った。従来方式では、相関係数計算での平均値を３に固定して計算した。閾値は０．３５－０．９０の間を０．０５刻みに実験した。提案方式では、予測順位から予測評価値への変換は線形補完で行なった。

5.3. 評価指標

・誤差二乗平均

予測能力を測る尺度として予測誤差、利用者が実際につけた評価値と予測評価値の差をEの二乗平均の平方根

$$\text{誤差} = \sqrt{E^2}$$

を用いて比較した。この指標は差の二乗項があることで、差の絶対値の平均より、予測が大きく外れたデータに敏感に反応する。

・予測可能率

従来のＳＩＦ方式では利用者間類似度の閾値のため評価予測できないことがあるが、評価予測出来た割合を予測可能率とした。

また、比較のため提案方式の予測能力を、従来方式で評価予測出来た情報と、出来なかった情報に分けて測ることもした。

5.4. 実験結果

従来方式と提案方式の予測誤差は表７のようになった。提案方式の予測誤差は従来方式の３分の１以下となった。

各閾値での性能を表８に示す。誤差 a は従来方式で評価予測出来た情報だけで測った提案方式の誤差、誤差 b は予測出来なかった情報で測った提案方式の誤差である。これを見ると、提案方式は常に従来方式より精度がよいことが分かる。また、従来方式は閾値を緩くしても予測精度の低下は起こしていない。これは、データを絞り込んでエントリー率を上げたためだと思われる。

表７：予測誤差

従来方式	提案方式
1.37 (閾値=0.35)	0.429

6. まとめ

ＳＩＦの改良として、インデックスの生成に従来の相関係数にかえて、多変量解析（双対尺度法）で生成した情報空間を使用する方式を提案した。また、実験により従来方式より評価予測能力が高いことが分かった。

現在、より多くのデータによる予測性能の評価、特に評価データ量と予測能力の関係の分析を行なっている。また最近、ニューラルネットと組み合わせた方式[7]や重み付き多数決法を使った改良[8]が提案されているため、それらとの比較も行なっていきたい。

表８：予測誤差（閾値別）

従来方式			提案方式	
閾値	誤差	予測率	誤差 a	誤差 b
0.35	1.37	0.915	0.437	0.327
0.40	1.39	0.854	0.444	0.332
0.45	1.43	0.748	0.458	0.331
0.50	1.48	0.572	0.475	0.360
0.55	1.52	0.445	0.489	0.374
0.60	1.54	0.354	0.511	0.377
0.65	1.57	0.266	0.529	0.386
0.70	1.54	0.207	0.576	0.382
0.75	1.60	0.135	0.636	0.387
0.80	1.64	0.102	0.665	0.393
0.85	1.56	0.081	0.707	0.395
0.90	1.52	0.060	0.761	0.399

参考文献

[1] "Special Issues on Information Filtering", Communication of the ACM, Dec 92, Vol.35, No.12, pp.26-81, 1992.

[2] "Experiences with GroupLens: Making Usenet useful again", Miller, B. et al., Proc. of 1997 Usenix Winter Technical Conference. pp.219-233. 1997.

[3] "Social Information Filtering: Algorithms for Automating "Word of Mouth"", Upendra Shardanand, Pattice Maes, Proc. of CHI'95, pp.210-217, 1995.

[4] "情報フィルタリングシステム", 森田, 速水, 情報処理 37(8), pp.751-758, 1996.

[5] "質的データの数量化", 西里, 朝倉書店, 1982.

[6] "情報の内容と他者の評価を利用した情報フィルタリング方式", 有吉, 市山, 電子情報通信学会第８回データ工学ワークショップ論文集, pp.49-54, 1997.

[7] "Learning collaborative information filters", Daniel Billsus, Michael Pazzani, Proc. of ICML'98, pp.46-54, 1998.

[8] "Collaborative Filtering using Weighted Majority Prediction Algorithms", Atsuyoshi Nakamura, Naoki Abe, Proc. of ICML'98, pp.395-403, 1998.