5M－8　特異値分解を用いたソーシャル情報フィルタリング方式

有吉 勇介
NECインターネットシステム研究所

1．はじめに

情報フィルタリングは，利用者の関心の学習と情報に対する評価を予測する技術であり，情報推薦サービスの基本技術である[1]．SIF(Social Information Filtering)方式は，他の利用者の評価に基いて評価予測を行なう方式である．キーワードや単語出現頻度などの情報の内容を利用するCBF(Content Based Filtering)方式では難しい，文芸的文書や画像や画像・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．

本稿では，このSIF方式の選別精度を改善する方式を提案する．従来のSIF方式では，1) 利用されていない評価情報が存在する，2) 情報の直交性を仮定している，3)評価値をそのまま使用している，といった問題点があった．これらの問題を解決するため，提案方式では特異値分解を導入し，各利用者ごとに評価値の変換を行う．さらに，提案方式が従来方式より選別精度が高いことを，技術文書推薦サービスのデータを用いた実験により確認した．

2．従来方式と問題

2.1　従来方式の構成

従来のSIF方式は利用者間類似度算出と評価予測の2段階から構成されている[2]．

①利用者間類似度算出：従来方式では各利用者毎に他の利用者との類似度リストを持っている．利用者間類似度は，両者の評価履歴から共に評価している情報に関する部分を抜き出し，その相関係数を使用する．

②評価予測：従来方式における情報aに対する評価の予測値は，情報aに他の利用者が与えた評価値を，利用者間類似度を重みとして重み付平均したものである．ただし，類似度が閾値以上の利用者のみを対象に重み付き平均をとる．

Social information filtering method using singular value decomposition.
Yusuke Ariyoshi
E-mail: y-ariyoshi@bx.jp.nec.com
Internet Systems Research Laboratories, NEC Corp.

2.2　問題点

従来方式は，以下の問題点が指摘されている．

1)未利用の評価情報：従来方式は類似度が閾値以下の利用者の評価は利用していない．また，共に評価している情報が無い利用者の評価も利用していない．

2)情報の直交性の仮定：利用者間類似度に用いている相関係数は，評価履歴ベクトルがなす角のコサインである．この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ている．しかし，この仮定は情報同士は似ているものも似ていないものもあるという現実と異なっている．

3)評価値をそのまま利用：評価値は，本来，順序の意味しか持たない．例えば，3つの情報(ア・イ・ウ)にそれぞれ評価値を1，2，3と付けた場合，ウイアの順で評価が高いことを表わすが，評価の数値の差に意味はない．アとイの差はわずかでイとウの差は非常に大きい利用者もいれば，その逆の利用者もいる．ところが，従来の評価値をそのまま使用して相関係数を計算するということは，各評価段階は利用者によらず等間隔であることを仮定している．

1)の問題により，実際には関心が類似しているのに共に評価している情報が無い利用者や，類似度が閾値以下の利用者にしか評価されていない情報は推薦できない．また，未利用の評価情報を活用できれば選別精度が向上することが考えられる．2)3)の不自然な仮定についても，現実に合うように変更できれば選別精度が改善できる可能性がある．

3．提案方式

提案方式では，問題点1)2)に対処するためにSIFを特異値分解によって生成される空間上で行なう．また，3)に対応するために評価値の尺度変換を行なう．

3.1　特異値分解を用いたSIF

提案方式では，次の性質を持つ情報空間によりフィルタリングを行なう．

その情報空間上では，情報は利用者による評価が似ているもの同士ほど近くなるように，利用者は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，また，利用者に高く評価された情報ほどその利用者の近くになるように配置する．次に，この情報空間上での利用者と情報との距離を測り，利用者の近くにある未評価の情報をその利用者に推薦する．

この情報空間での利用者と情報の位置は，評価情報を格納した行列を特異値分解[3]することで算出することができる．利用者と情報の距離は，それぞれの位置を表わす座標ベクトルの相関係数で求められる．

提案方式では，全ての評価情報を用いて情報空間を生成する．そのため，類似度が閾値以下の利用者や共に評価している情報が無い利用者にしか評価されていない情報であっても，利用者に推薦することができ，1)の問題が解決する．また，この情報空間上では，情報は互いの類似性に応じて配置されるので，2)の問題も解決する．

3.2 尺度変換

提案方式では，3)の問題を解決するために評価値を順位に変換する尺度変換を行なう．尺度変換では利用者毎に順位付けと同点処理を行う．順位付け処理は，評価済み情報に順位付けを行い，同点処理では異なる情報に同じ評価値が付いている場合，可能性のある全ての順位を平均した順位を求める．提案方式では，フィルタリング処理を行なう前に尺度変換を行なう．

評価値のままでは，各評価段階は，利用者間でも利用者内でも等間隔であった．しかし，順位に変換することで，評価段階の間隔は各利用者の評価付けの特徴に応じて異なるものになり，3)の問題が解決できる．

4．評価実験

実験では，社内で行った技術情報推薦サービスのデータを用いて，方式毎の選別精度を測定した．データは，推薦文書に対し利用者が関心に応じて5段階評価したものであり，そのうちある程度継続利用した 89 人のデータを実験で使用した．使用した評価データは 12,200 件で，文書数は 2,149 文書である．選別精度の測定は，評価データを 5 ブロックに分割し，1つのブロックを残り 4 ブロックのデータから予測することを，全ブロックについて行なった．

実験結果：図の横軸はカットオフ順位，縦軸はその順位以上を推薦した時の適合率である．適合率は推薦情報のうち実際の評価が4以上のものの割合である．データ中に実評価が 5 のものが 871 件，4 以上が 2,270 件あるので，実際のサービスでのカットオフ順位は 870～2,300 辺りになると予想される．その範囲の適合率を比較すると，提案方式が従来方式より良いことが分かる．

5．おわりに

本稿では SIF の改良として，特異値分解により生成される情報空間を用いてフィルタリングを行なう方式を提案し，従来方式より評価予測能力が高いことを実験的に確かめた．情報内容に基づく方式を特異値分解を用いて改良したものは LSI と呼ばれる[3]．本稿の方式は LSI の SIF 版と言うこともできる．今後，本方式を現在開発中の推薦エンジン[4]に組込み，実システムでの評価を行ないたい．

参考文献

[1] "Special Section: Recommender Systems", CACM, Vol.40, No.3, pp.56-89, Mar 1997.

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

[3] S.T.Dumais, "Using LSI for information filtering: TREC-3 experiments", Proc. of TREC-3, pp.219-230, 1995.

[4] 福島，有吉，"レコメンデーションエンジン KnowledgeRadar/R の開発",情報処理学会第 60 回全国大会講演論文集(3), pp.127-128, 2000.

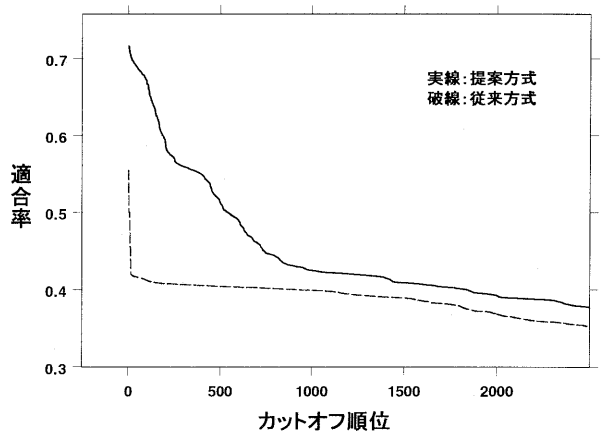

図 1：適合率－カットオフ順位