組合せ情報フィルタ方式の信頼度を用いた精度改良†

4P－9

有吉 勇介　福島 俊一
NECヒューマンメディア研究所

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評価の予測を行う技術であり，情報推薦サービスの基本技術である[1]．評価予測の方式には，キーワードや単語出現頻度などの情報の内容に基づくＣＢＦ(Content Based Filtering)と，情報に対する他の利用者の評価に基づくＳＩＦ(Social Information Filtering)の２種類の方式があり，両者の組合せ方式も研究されている[2,3]．本稿では評価予測の精度を高めるために，予測評価値の信頼度を考慮して組合せる方式を提案する．

2.従来方式

ＣＢＦ：ＣＢＦでは，利用者の情報に対する評価を分析し，評価の高い情報に含まれる単語の重要度を上げ，評価の低い情報に含まれる単語の重要度を下げることで，単語の重要度を学習する．評価予測は，学習した単語の重要度と情報中の単語の出現頻度を照合することで行う．

ＳＩＦ：ＳＩＦでは，まず推薦しようとしている利用者とそれ以外の利用者との評価履歴を比較することで興味の似た利用者を見つけ出す．そして，興味が似た利用者が高く評価している情報を推薦する．

組合せ方式：ＳＩＦは誰もまだ評価していない新規情報を推薦することはできない．一方ＣＢＦは，図表等の単語頻度に表れない情報の価値を推薦に織り込むことが難しい．さらに，その利用者が評価済みの情報に表れない単語を多く含む情報に対してはフィルタリング精度が低くなってしまう．そこで従来の組合せ方式では，両方式の特徴を生かした推薦を行うために，ＳＩＦとＣＢＦを予測評価統合と学習強化の２個所で組合せている．**予測評価統合**ではＣＢＦとＳＩＦの予測評価結果を統合する．まだ誰も評価していない新規情報はＣＢＦの予測評価値を用い，他の利用者が評価済みの情報はＳＩＦの予測評価値を用いる．**学習強化**では，ＣＢＦの単語重要度の学習をＳＩＦ結果で強化する．単語重要度を，その利用者の評価済み情報に加えて，ＳＩＦによる予測評価値を評価値の代わりにすることでその利用者が未評価の情報からも学習する．

3.信頼度を用いた改良組合せ方式

提案方式は従来の組合せ方式の精度を高めるために予測評価値の信頼性を導入する．従来の予測評価統合では新規情報かどうかでＣＢＦとＳＩＦを切り替えていたが，評価済み情報であっても評価者が少人数の場合はＣＢＦのほうが信頼性が高いであろう．学習強化についても，ＳＩＦの予測評価のうち信頼性の高いものを用いたほうが，フィルタリング精度が高くなると考えられる．

予測評価値の信頼性は，予測に用いたデータの量（例えばＣＢＦではプロファイルの大きさ，ＳＩＦでは類似利用者の人数など）が多いほど高い．そこで提案方式ではデータ量から予測評価値の信頼性を推定する．

4.比較実験

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利用した．サービスでは推薦文書に対し利用者は関心に応じて５段階評価する．実験では，ある程度継続利用した４５人のデータを使用した．評価データは１０７０９件で，文書数は３０００文書である．

誤差推定方式：今回の実験では，信頼性の指標として誤差（予測評価値と実評価値の差の二乗）を用いた．また，誤差の推定はデータ量の多項式によって行った．データ量としてはＣＢＦでは重要度学習済みの単語種数と情報

† Improvement of Combination Information Filtering Methods based on Reliabilities

Yusuke Ariyoshi, Toshikazu Fukushima
E-mail:{ariyoshi,fuku}@hml.cl.nec.co.jp
Human Media Research Laboratories, NEC corp.

本研究の一部は日本情報処理開発協会の次世代電子図書館システム研究開発事業の一環として行われている．

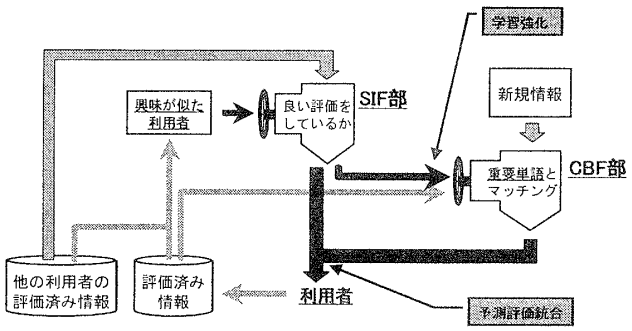

図 1：組合せ方式の構成

に含まれる単語のうち重要度学習済みの単語種数を，ＳＩＦは類似利用者全体での利用者間類似度の総和を用いた（これらは予備実験で誤差推定精度が高いことが分かったためデータ量として用いた）．

実験では上述の評価データを用いて 10 Cross Validation を行い，学習強化と予測評価統合について比較実験を行った．

学習強化：次の２種類の方式を比較した．ただしNは利用者が実際に評価した情報の件数である．**単純強化**：ＳＩＦの予測評価値の大きいものをN／２件，小さいものをN／２件学習に追加する．**誤差考慮強化**：ＳＩＦの予測評価値の大きいものをN件，小さいものをN件選び，その２N件から推定誤差が小さいN件を学習に追加する．

予測評価統合：次の３種類の方式を比較した．**単純統合**：推定誤差を考慮せず，ＳＩＦで予測可能なものはＳＩＦ予測評価値を，予測出来ない場合はＣＢＦ予測評価値を利用する．**切替え統合**：推定誤差の小さい方を選択する．**ブレンド統合**：ＣＢＦとＳＩＦによる予測評価値を推定誤差の比で内分した値を最終的な予測評価値にする．

実験結果：図２はＣＢＦ単独で学習強化方式を変更した結果であり，図３は統合方式を比較した結果である（学習強化方式は誤差考慮強化方式に固定）．横軸はカットオフ順位，縦軸は適合率である．適合率は推薦情報中の実評価が４以上のものの割合である．データ中に実評価値が５のものが７３０件，４以上が２０００件あるので，実際のサービスでのカットオフ順位は７００～２０００辺りになると予想される．その範囲で適合率を比較すると，学習強化方式は誤差考慮強化方式，統合方式は切替え方式の適合率が良いことが分かる．

5.おわりに

本稿ではＳＩＦとＣＢＦの組合せ方式について，予測に使用した評価データ数や単語数などから予測評価の誤差を推定し，それにより組合せ方を変える改良を提案した．比較実験の結果，提案方式は従来方式より適合率が高かった．また，実験に用いたデータは４５人分と少なかったが，逆にＳＩＦが有効でない小人数の利用者に対しても提案方式は適合率を改善する効果があるといえる．

参考文献

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

[2] 有吉，市山，"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",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.

[3] M. Balabanovic, Y. Shoham, "Fab: Content Based Collaborative Recommendation", CACM, Vol.40 No.3, pp.66-72, Mar 1997.

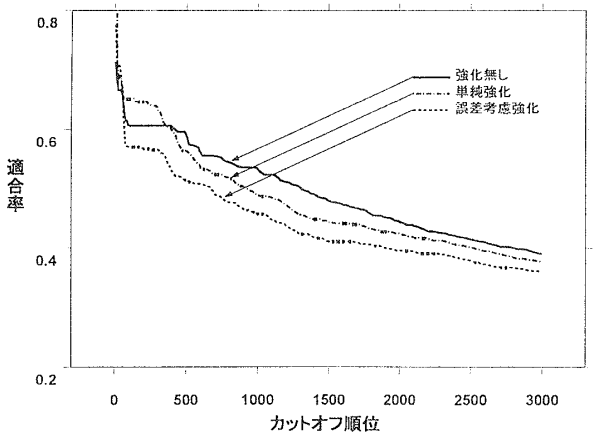

図 2：学習強化方式比較

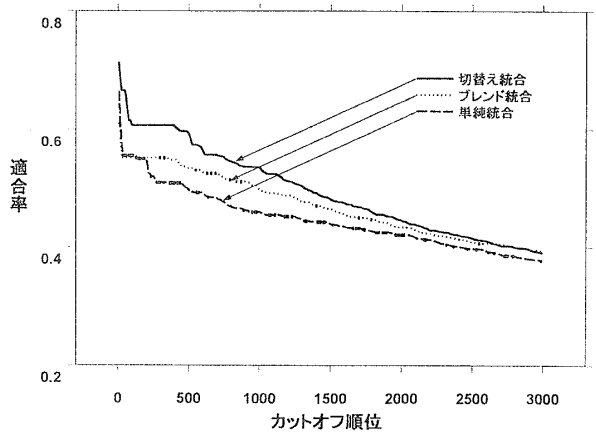

図 3：予測評価統合方式比較