3J1-NFC1a-1

確率モデルを用いたTV視聴データの分析

Analysis of TV Audience using Probabilistic Model

村上 知子　　中田 康太
Tomoko Murakami　Kouta Nakata

株式会社東芝 研究開発センター
Corporate R&D Center, Toshiba corporation

One of the issues of recommender system is cold-start, which is the problem that recommender system does not work well under the situation that the user has just started using it. Although it is a common problem in both content-based and collaborative methods, it is rather serious in content-based method because of the lack of information. In this paper, we address cold-start in TV recommender system using content-based method. We propose a method based on cluster model and content data to avoid cold-start and verify its effectiveness through TV show recommendation.

1. はじめに

インターネットの普及とその商業利用の拡大により，インターネットの利用者がアクセスすることが可能な電子化された情報 (コンテンツ) の量は飛躍的に高まった．その結果として，システムが利用者の行動をモニタリングし，利用者が必要としているであろうコンテンツを予測して利用者に提示する推薦システムに注目が集まり，様々なシステムが実用化されている．推薦システムの運用時の課題のひとつに，推薦システムの利用を開始して間もない利用者に対して妥当な推薦が困難であるという *cold-start* 問題が挙げられる [Herlocker 04, Breese 98]．推薦システムは利用者の行動履歴データに基づいて利用者の嗜好を予測するため，行動履歴データが少ない推薦システムの利用初期の段階においては，利用者の嗜好を的確に予測することが困難である．そのため，大量のコンテンツの中から嗜好にかなうコンテンツを自動的に発見することが可能な推薦システムの恩恵を享受しないうちに推薦システムの利用を止める利用者が多く存在する．

本論文では，利用者の選好情報が十分に得られない推薦システムの利用開始時に，早期に的確に利用者の嗜好にかなうコンテンツを推薦するための手法を提案する．提案手法では，まず多くの利用者のプロファイル情報や選好情報をもとにあらかじめ利用者のクラスタモデルを作成し，その後，推薦対象利用者と類似するクラスタモデルと行動履歴データに基づいて推薦アイテムを決定する．

本論文は以下のように構成される．2 章でテレビの放送番組推薦システムとそこでの *cold-start* 問題に関して説明する．3 章で利用の初期段階においても早く的確に利用者の嗜好にかなうコンテンツを推薦するための手法を述べる．4 章でテレビ番組の推薦における評価実験に関して説明する．5 章で結論と今後の展望を述べる．

連絡先: 村上 知子，(株) 東芝研究開発センター，
住所:〒 212-8582 川崎市幸区小向東芝町 1
電話番号:044-549-2406
メイルアドレス:tomoko.murakami@toshiba.co.jp

2. 課題

衛星放送の普及と全国的な地上波デジタル放送の開始により，本格的な多チャンネル時代が到来している．2009 年 1 月時点において，1 日あたりに国内で放送される番組数は約 3,000 件に達し，そのうち主要 7 放送局による番組数はわずか 1 割に過ぎず，それ以外は専門チャンネルによる内容が特化された放送番組である．一方，利用者にとって放送コンテンツを知る術も，従来の新聞のテレビ欄に代表される書誌上の番組表から電子番組表 (EPG) へデジタル化が進んだ．しかし，それらのコンテンツを視聴する機器の表示スペースに対する制約と，時間と放送局をそれぞれ縦横軸としたインタフェースは旧態依然であるため，大量の放送コンテンツの中から利用者が自分の嗜好にかなうコンテンツを発見するのは容易な作業ではない．このように，放送環境の整備に伴う放送コンテンツの多様化・大量化とともに，放送コンテンツの推薦システムへの期待が高まっている [村上 07]．

テレビ放送番組の推薦システムに関しては，協調フィルタリング [Resnick 94] を用いた研究事例が過去に紹介されている [Smyth 00, Ali 01]．他の利用者との類似性に基づく推薦手法である協調フィルタリングは，新たに視聴を開始した利用者と類似する視聴傾向を持つ他者の特定が困難であるため，利用開始初期段階では良い推薦が期待できない．また，テレビ放送においては，番組は毎回常に変化していることが多いため，コンテンツに対する評価情報 (rating) を維持・管理するためのコストが高く，番組放送前にそれらの情報を得られないという特徴がある．そのため，テレビ放送番組の推薦には協調フィルタリングではなく，コンテンツに基づく推薦が適していると考えられる．しかし，コンテンツに基づく推薦においても，利用者の視聴履歴情報が乏しいために新たに利用を開始した利用者に対して妥当な推薦が困難である．活用可能な情報が利用者自身の視聴履歴情報に限られているため，cold-start は協調フィルタリングよりもむしろコンテンツに基づく推薦においてより深刻な問題であるとも指摘されている [神嶌 07]．

3. 提案手法

本論文では，利用者の選好情報が十分に得られない推薦システムの利用開始時に，早期に的確に利用者の嗜好にかなうコンテンツを推薦するための手法を提案する．提案手法では，

まず多くの利用者のプロファイル情報や選好情報をもとにあらかじめ利用者のクラスタモデルを作成する．ここで，プロファイル情報とは，利用者の年齢や性別などのデモグラフィック情報や利用者の嗜好を表わすキーワードなどの情報を指す．そして，推薦対象利用者と類似するクラスタモデルと行動履歴データに基づいて推薦アイテムを決定する．

一般に，コンテンツに基づく推薦では，利用者の嗜好に関するデータを用いて嗜好モデルを学習，それに基づいて未知のコンテンツに対する選好度[*1]を予測する．そして，M 件から成る全コンテンツ集合から選好度の上位 N 件のコンテンツの情報を利用者に提示する．視聴履歴情報 $\{R_a\}$ に基づいて，ある利用者 U_a があるアイテム I を好む (Y) 確率 (選好度) の計算は以下のように定式化される．

$$Pr(Y|I) = Pr(Y|I, \{R_a\}) \tag{1}$$

一方で，クラスタに基づく推薦では，多くの利用者 U のプロファイル情報や選好情報をもとにクラスタモデル C を作成し，それに基づいて選好度を計算する．U_a と類似するクラスタが C_a である時，クラスタに基づいて U_a の I に対する選好度の計算は以下のように定式化される．

$$Pr(Y|I) = Pr(Y|I, C_a) \tag{2}$$

U_a と類似するクラスタ C_a は，利用者プロファイル情報や行動履歴情報などから以下のような推定される．

$$C_a = maxPr(C_j|D) \tag{3}$$
$$C_a = maxPr(C_j|D, \{R_a\}) \tag{4}$$

ここで，D は利用者プロファイル情報を指す．式 (3) は D に基づいて U_a の類似クラスタ C_a を決定する方法で，U_a の行動履歴情報が十分に得られない利用開始時にも適用可能である．式 (4) は D と $\{R_a\}$ に基づいて C_a を決定する方法で，行動履歴情報がある程度得られる状況において適用される．

上記のアプローチに対して提案手法では，クラスタと行動履歴情報の両方を用いて U_a の I に対する選好度を以下のように求める．

$$Pr(Y|I) = Pr(Y|I, \{R_a\}, C_a) \tag{5}$$

U_a の行動履歴情報が十分に得られない利用開始時には利用者のクラスタに基づく予測の効果が高いものの，行動履歴情報がある程度得られるようになるとコンテンツに基づく予測の方が優位になると考えた．

4. 実験

4.1 データ

実験では，地上波放送の番組を対象として 2008 年 2 月 7 日から 20 日までの 2 週間の番組に関する情報を収集した．放送番組に関する情報として，放送形態 (地上アナログ/デジタル，BS/CS 放送の識別)，番組識別 (放送日ごとにユニークな番組 ID)，放送日時 (開始時刻，放送時間)，放送チャンネル，番組タイトル，出演者，番組ジャンル，番組内容などの情報が公開されている．ただし，すべての番組にそれらの情報が付与されているわけではない．

視聴履歴データは，アンケート調査を通じて網羅的に収集した．被験者の嗜好を把握するため，放送後にすべての番組コンテンツを対象に，(1) 実際に視聴・録画した番組，(2) 実際には視聴・録画していないが視聴したかった番組，という 2 つの項目に対する回答を収集した．アンケート調査の結果，表 1 に示すように，約 250 人の被験者の 2 週間にわたる視聴履歴データが得られた．

表 1: データセット

被験者数	250 人
プロファイル情報	性別，年代
データ収集期間	2008/2/7-20
番組数	9430 件
視聴番組数 (日・人平均)	9.4 件

表 2: プロファイル情報に基づくクラスタ

クラスタ	サイズ (単位:人)
20-24 歳の男性	23
25-29 歳の男性	3
30 歳代の男性	26
40 歳代の男性	25
50 歳代の男性	24
60 歳代の男性	23
20-24 歳の女性	22
25-29 歳の女性	2
30 歳代の女性	26
40 歳代の女性	25
50 歳代の女性	26
60 歳代の女性	26

4.2 評価

実験では，新しい利用者が推薦システムの利用を開始した際の，式 (1) で表されるコンテンツに基づく推薦手法 (CB)，式 (2) で表されるクラスタモデルに基づく推薦手法 (CL) や式 (5) で表されるクラスタモデルとコンテンツに基づく推薦手法 (CBCL) によるシステムの振る舞いを検証した．

コンテンツに基づく推薦手法 (CB) の適用においては，利用者が推薦システムの利用を開始して間もない状況を再現し，実験を行った．具体的には，収集した番組および表 1 に示す視聴履歴データを放送時間順にソートし，利用者ごとに学習データ内の視聴番組数 $Rnum$ が 1 件から 50 件になるようにある時刻でデータを学習データとテストデータに 2 分割した．そのようにして得られた 50 件の学習・テストデータを用いて，視聴アイテムが 1 件ずつ増加する場合の番組視聴の予測精度を求めた．実験の結果は，テストデータ中の視聴アイテムに対する選好度 ($Pr(Y|I_{pos})$) と学習器による選好度上位のアイテムに占める実際の視聴アイテムの割合を表す適合率 ($Precision$) を用いて評価した．学習器には，番組の出演者情報から視聴を予測する naive Bayes モデルを利用した．

クラスタモデルに基づく推薦手法 (CL) の適用においては，性別と年代から成る利用者のプロファイル情報に基づいてクラスタを定義した．各クラスタに該当する利用者の視聴履歴データを学習データとしてクラスタモデルを作成し，利用者の視聴履歴データをすべてテストデータとした．そして，式 (3) に従いプロファイル情報から類似するクラスタ C_a を定め，式 (2) に従いアイテムに対する選好度を求めた．表 2 は各クラスタとそれを類似クラスタとする利用者数を示している．

*1 選好度が 2 値ならば { 好き, それ以外 } を，選好度が多値ならば嗜好の程度を指す

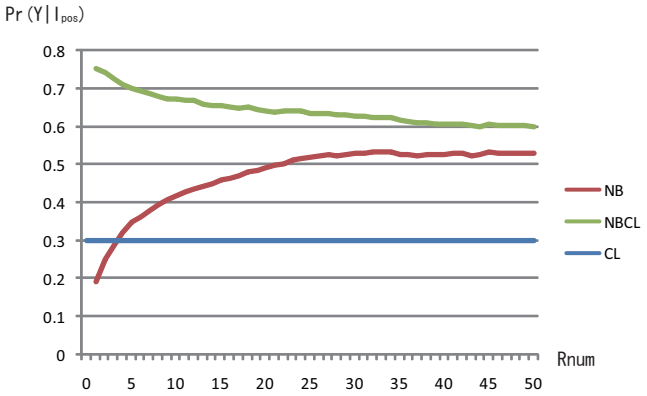

図 1: 視聴アイテムに対する選好度の推移

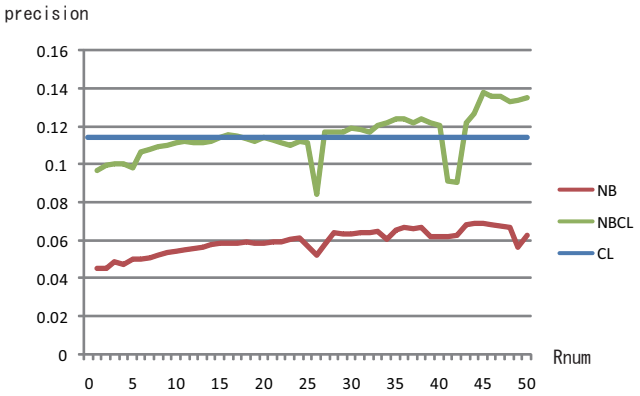

図 2: 推薦の適合率の推移

クラスタモデルとコンテンツに基づく推薦手法 (CBCL) の適用においては，コンテンツに基づく推薦手法 (CB) 適用時と同様に，利用者が推薦システムの利利用者が推薦システムの利用を開始して間もない状況を再現し，実験を行った．利用者プロファイル情報から特定された類似クラスタに該当する利用者の視聴履歴データと推薦対象利用者の視聴履歴データを利用してモデルを作成する．その際，推薦対象利用者の視聴履歴データを放送時間順にソートし，学習データ内の視聴番組数 $Rnum$ が 1 件から 50 件になるように学習データとテストデータに 2 分割した．そのようにして得られた 50 件の学習・テストデータを用いて，視聴アイテムが 1 件ずつ増加する場合の番組視聴の予測精度を求めた．

図 1 と図 2 に結果を示す．図 1 は学習データにおける視聴番組数 $Rnum$ と視聴番組に対する選好度 $Pr(Y|I_{pos})$ の関係を表している．NB は $Rnum \leq 4$ のときには低調であるが，$Rnum \leq 25$ の場合に学習データ内視聴番組数が増加するにつれて $Pr(Y|I_{pos})$ が増加している．視聴番組数が少ない状況では学習器が確からしい選好度を計算できていないことが分かる．一方，NBCL は $Rnum$ の小さいときから高い値を示すが徐々に低下することが読み取れる．NBCL は，類似クラスタに該当する利用者の視聴番組データと推薦対象利用者の視聴履歴データを用いて学習データを作成するため，利用開始時には学習データにおける視聴番組の割合が多いことから $Pr(Y|I_{pos})$ が高い値を示している．図 2 は学習データにおける視聴番組数 $Rnum$ と推薦における適合率 $precision$ の関係を表している．NB と NBCL 共に，学習データ内視聴アイテム数が増加するにつれて $precision$ が増加する傾向にあること，NBCL は NB より常に $precision$ が高く優位に推移していることが読み取れる．CL では $precision = 0.1140$ となり，$Rnum \leq 10$ であるような推薦システムの利用開始時には最も高い精度が得られていることが分かる．本結果から，利用者の視聴データ数が少ない推薦システムの利用開始時には学習器による予測精度が向上しない cold-start 問題が TV 番組の推薦においても確認された．また，クラスタモデルに基づく推薦手法は推薦システムの利用開始時に有効であるが，利用開始後視聴アイテム数がある程度得られれば NBCL が精度を凌ぐことが明らかになった．

5. おわりに

本論文では，推薦システムの利用の初期段階においても早く的確に利用者の嗜好にかなうコンテンツを推薦する手法を提案した．クラスタモデルに基づく推薦手法やクラスタモデルとコンテンツに基づく推薦手法は，行動履歴情報が十分に得られない利用開始時に有効であることが分かった．今後は，プロファイル情報に加えて行動履歴情報も利用した類似クラスタ推定 (式 (4)) の効果やクラスタモデルを導入した協調フィルタリングの予測精度などの cold-start 改善のための手法を検証していく．また，本論文では対象としなかったが，cold-start のもう一方の課題である推薦対象として新たに導入されたアイテムを推薦する困難さに対しても解決案を考えていきたい．

参考文献

[神嶌 07] 神嶌 敏弘: ”推薦システムのアルゴリズム (1)”，人工知能学会誌, Vol.22, No.6, pp.826-837, (2007).

[村上 07] 村上 知子: ”AV 機器利用者に対する放送コンテンツの推薦” Vol.48, No.9, pp.984-988, (2007).

[Ali 01] K. Ali and W. V. Stam: TiVo: Making Show Recommendations Using a Distributed Collaborative Filtering Architecture, In proc. of the International Conference on Data Mining and Knowledge Discovery (ACM SIGKDD), pp.408-413, (2001).

[Smyth 00] B. Smyth and P. Cotter: A Personalized Television Listing Service, Communications of the ACM, Vol.43, no.8, pp.107-111, (2000).

[Breese 98] J. Breese, J. Herlocker and C. Kadie.: Empirical analysis of predictive algorithms for collaborative filtering, In proc. of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98), pp.43-52, (1998).

[Resnick 94] P. Resnick, N. Iacovou., M. Suchak, P. Bergstorm. and J. Riedl: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, In proc. of International Conference on Computer Supported Cooperative Work (CSCW), pp.175-186, (1994).

[Herlocker 04] J. Herlocker, J. Konstan, L. Terveen and J. Riedl: Evaluating Collaborative Filtering Recommender Systems, J. of ACM Transactions on Information Systems, Vol.22, No.1, pp.5-53, (2004).