3H2-03

帰納学習を用いた映画推薦システム

A Movie Recommender System Based on Inductive Learning Algorithms

李 鵬[*1]
PENG LI

山田 誠二[*2]
SEIJI YAMADA

[*1] 東京工業大学
Tokyo Institute of Technology

[*2] 国立情報学研究所
National Institute of Informatics

The tremendous growth in the amount of available information and the number of visitors to Web sites in recent years poses some key challenges for recommender systems. New recommender system technologies are needed that can quickly produce high quality recommendations, even for very large-scale problems. In this paper we propose a new recommender system technology based on inductive learning. To inspect the effectiveness of this technology, we set up a movie recommender system based on inductive learning and make online experiments for the evaluation of our system. Our results suggest that inductive-learning-based technology is available for the solution of the challenges for collaborative filtering based systems, while at the same time providing high performance recommendations.

1. はじめに

近年，ニュース，音楽，動画などの様々なコンテンツが提供されているが，膨大な量のコンテンツの中からユーザが自分の嗜好に合ったコンテンツを獲得することは困難になっている．このような状況を解決するための一つのアプローチとして，推薦システム(Recommender System) [1]が開発されている．推薦システムの要素技術である協調フィルタリング[2]は研究，実用ともに広い範囲で成功をおさめているが，ユーザ数と情報量の急増に伴って，スパーシティの問題による推薦精度の低下と，スケーラビリティの問題による推薦スループットの低下といった問題を抱えている．

本研究では，これらの問題の解決を試みるに当たって，帰納学習(Inductive Learning)を用いた推薦手法を提案する．帰納学習を用いた推薦手法では，ユーザ間で共有できるアイテムのコンテンツプレファレンスを推薦に用いることでスパーシティの問題を解決する．また，学習時の計算コストは，実際的に十分に小さいものであり，瞬時にユーザプロファイルを生成できる．さらに，スケーラビリティの増大に伴って，推薦時の計算コストが膨大になることもない．帰納学習を用いた推薦手法の有効性を調べるために，映画推薦システムを実装し，評価実験を行った．

2. 既存の情報推薦手法と問題点

2.1 内容に基づくフィルタリング

内容に基づくフィルタリングとは，ユーザの興味や関心を記述したユーザプロファイルに基づいて，情報源から次々と送られてくる情報からユーザの関心のあるものを提示したり，優先度を与えたりする手法である．

内容に基づくフィルタリングにおいて，システムを利用しているユーザが複数存在したとしても，個々のユーザの情報収集は独立していて情報の共有がされていないため，効率的な情報収集が困難である．同じことに興味を持った他のユーザが存在したとしても互いに情報共有できないため，新しい情報にユーザプロファイルと適合するデータが含まれていなければ，実際には興味を持つ情報であっても推薦されないことがある．そのため，新しい情報で意外性のある情報は推薦されないという問題がある．意外性がある情報とは，ユーザがそれまで見たことがなく，ユーザプロファイルに関連する特徴量を含まないものである．

2.2 協調フィルタリング

内容に基づくフィルタリングにおいては意外性のある情報を発見することは困難である．単一のユーザプロファイルにおいて，適合率を上げつつ再現率を上げるには，ユーザのプロファイルに含まれていない情報を他から集める必要がある．そのための情報として他のユーザの情報が考えられる．他の複数ユーザのプロファイル情報を用いることにより，ユーザプロファイルに含まれない部分の情報を見つけることが可能となる．このようなユーザ間の協調を支援する手法として協調フィルタリングがある．

協調フィルタリングとは，ユーザの情報収集行動から興味・関心・意図などのユーザの情報に対する取捨選択の基準を収集し，類似した選択基準を持つユーザに提供することにより情報収集を支援するための手法である．

協調フィルタリング技術は研究，実用ともに広い範囲で成功を収めているが，近年のユーザ数と情報量のすさまじい増加に対して，以下の問題を抱えている．

● スパーシティ（Sparsity）

スパーシティは，ユーザが評価したアイテムの数に対して，アイテムの総数があまりにも大きい場合を指す．多くの商用推薦システムは大規模なアイテムセット（例えば，Amazon.com は本と映画を推薦し，CDnow.com は音楽アルバムを推薦する）を使用する．それらのシステムでは，ユーザはアイテム全体の 1%以下しか購入しない[3]．従って，Nearest Neighbor アルゴリズムに基づく推薦システムは，特定のユーザに対してアイテムの推薦を行えない可能性もある．結果として推薦の精度に欠けることになる．

● スケーラビリティ（Scalability）

Nearest Neighbor アルゴリズムでは，ユーザとアイテムの増加に従って，膨大な計算が必要になる．1,000,000 単位のユーザ，並びに，アイテムに対し，現存するアルゴリズムで稼動している実用的なオンライン推薦システムは，深刻なスケーラビリティの問題を抱えている[3]．

連絡先 氏名：李 鵬，所属：東京工業大学(NII)，住所：国立情報学研究所 101-8430 東京都千代田区一ツ橋 2-1-2
メールアドレス：li@ntt.dis.titech.ac.jp

●トランスペレンシー(transparency)
ユーザに推薦される情報はユーザの嗜好との関連性が明確にされず，どのような経緯でユーザに推薦されたのか分かりにくい．

3. 帰納学習を用いた推薦手法

3.1 帰納学習と決定木

帰納学習とは，一般的な規則の具体例から規則を推論する学習手法で，人工知能では大きな研究分野となっている．その中で，分類手法としての決定木[4]は，歴史も古く計算時間も比較的短いため，現在ではデータマイニングの標準的な技法の一つとなっている．

決定木とは，データ項目間の関係を木構造で表示する分析手法である．決定木はノードとリンクから構成され，各ノードには分類する属性，ノードとその下位ノードとを結ぶリンクには属性値がそれぞれ対応付けされている．ただし，下位ノードは最上位ノードからのリンク属性値により分類されたクラスを表現する．決定木は分類型の知識を表現するのに適している．

決定木を生成する主な学習アルゴリズは[Breiman,1984]らによる CART や[Quinlan,1986，1993]による ID3，C4.5[5]，C5.0 がある．本論文では，実用的なデータマイニング法としての評価が確立されている C4.5 を使用する．

3.2 C4.5 による推薦

帰納学習を用いた推薦手法の基本的な考え方は，アイテムに対しての離散的評価値とアイテムの属性であるコンテンツプレファレンスをそれぞれ決定木のクラスと属性に対応させ，分類を行うことである．帰納学習を用いた推薦過程の詳細を説明する．

①トレーニングデータの入力：インタフェースを通してユーザにいくつかのアイテムに対する評価を付けてもらう．

② 決定木の構築：アイテムに対する評価とそのアイテムのコンテンツプレファレンスを用いて決定木を構築する，構築された決定木をユーザプロファイルとして使用する．

③ 未評価アイテムの分類：ユーザの嗜好を表した決定木を使ってユーザがまだ評価してないアイテムを分類する．分類されたクラスはアイテムに対する予測評価値と対応している．

④ 推薦候補リストの作成：未評価アイテムを予測評価値の高い順に並べて，推薦候補リストを作成する．

⑤ 推薦結果の提示：推薦候補リストを，インタフェースを通してユーザに提示する．

⑥ トレーニングデータの追加：提示されたアイテムに対してユーザがさらに評価した場合，そのアイテムをトレーニングデータに追加して，次回の推薦における決定木の構築に用いられる．決定木はインクリメンタルに生長してより精密になり，推薦精度が向上する．

3.3 協調フィルタリングの問題点の克服

協調フィルタリング手法を用いた推薦システムでは，ユーザ数と情報量のすさまじい増加に伴って，データの希薄性の問題による推薦精度の低下と，スケーラビリティの問題による推薦スループットの低下などの問題を抱えている．これらの問題に対して，帰納学習を用いた手法を以下のように解決を試みている．

● スパーシティ（Sparsity）
機能学習を用いた推薦手法では，ユーザ間の関係ではなく，アイテムの属性であるコンテンツプレファレンスに注目する．コンテンツプレファレンスは管理者かユーザに予め付けられていて，ユーザ間で共有される．コンテンツプレファレンスを持つアイテムであれば，決定木によって予測評価値を計算することができるので，データがスパースになることはない．

● スケーラビリティ（Scalability）
C4.5 学習アルゴリズムでは，計算量はトレーニングデータの数に依存する．m をトレーニングデータの数として，$O(m^2)$ の計算量を必要とする．実際には一人のユーザが評価するアイテムの数はせいぜい数十個なので，瞬時に決定木を生成できる．極まれにトレーニングデータの数が 1000 を超えるような場合は，選択的サンプリング手法[6]を用いる．全てのデータを学習に用いるのではなく，選択的にデータをサンプリングし，メモリに載せて学習を行う手法である．例えば，文献[6]では，既に選ばれたデータを複数回リサンプリングして，そこで得られたデータセットから複数の決定木を生成し，これらを用いてクラスを予測したときの予測値が最も割れるようなデータを選択的にサンプリングすることを繰り返す方法をとっている．この方法ではリサンプリングの繰り返しに計算時間がかかるものの，それは高々サンプル数の線形オーダであり，サンプル数の二乗に比例して計算時間がかかる決定木生成部分においてサンプル数を劇的に減らしているので，トータルとして高い Scalability を達成することができる．しかも，分類予測精度は全データを用いたものとほとんど変わらないといった有効性を持っている．

また，推薦時の計算量はアイテムの総数に依存する．アイテムの総数をnとすると，計算量は となる．従って，ユーザ数とデータ量が増加しても，計算量が膨大になることはない．

●トランスペレンシー(transparency)
高い Readability(知識の読みやすさ)は決定木の利点の一つである．決定木を用いた推薦手法では，推薦される情報はユーザの嗜好との関連性が明確にされているだけではなく，決定木の構造自体がユーザの嗜好を表しているので，嗜好に関する詳しい情報をユーザに提示することができる．

4. 帰納学習を用いた映画推薦システム

3章で提出した帰納学習の推薦手法を用いた映画推薦システムを実装した．オンラインで実験が行えるために WWW に公開している．URL は http://liorlee.ddo.jp ，実装言語は ruby1.8.1，WEB サーバは apache1.3.28 を使用した．

4.1 コンテンツプレファレンスとクレジットプレファレンス

本システムで使用する映画データはコンテンツプレファレンスとクレジットプレファレンスの 2 種類ある．コンテンツプレファレンスはタイトルごとにストーリ展開の面白さ,コメディ,アクション,感動,ほのぼの,ホラー,演技の 7 つの属性を持っている．属性値は 1 から 5 までの整数値を取り，値が高いほど優れている．クレジットプレファレンスはジャンル，監督，出演者といった情報である．

4.2 システムの構成と推薦の流れ

本システムの構成を図 1 に示す．ユーザは GUI を通していくつかの映画を評価し，それをトレーニングデータとして C4.5 に与える．C4.5 によりトレーニングデータxを用いて，決定木 を作成する．決定木を用いてユーザがまだ評価してないアイテムを分類し，推薦候補リストを生成する．推薦候補リストを受け取った評価エンジンは，ユーザが指定したクレジットプレファレンスを分析し，推薦候補リストのソートを行う．最後に，評価エンジンは推薦結果を，GUI を通してユーザに提示する．

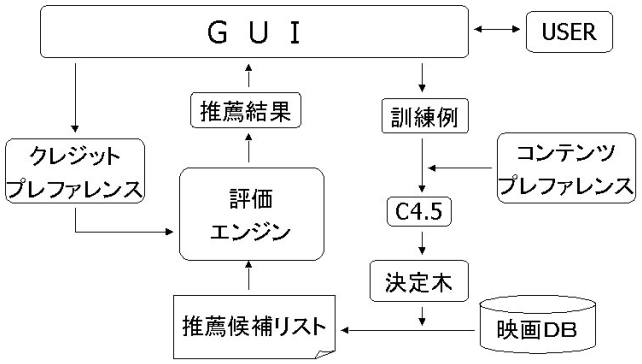

図 1 システム構成図

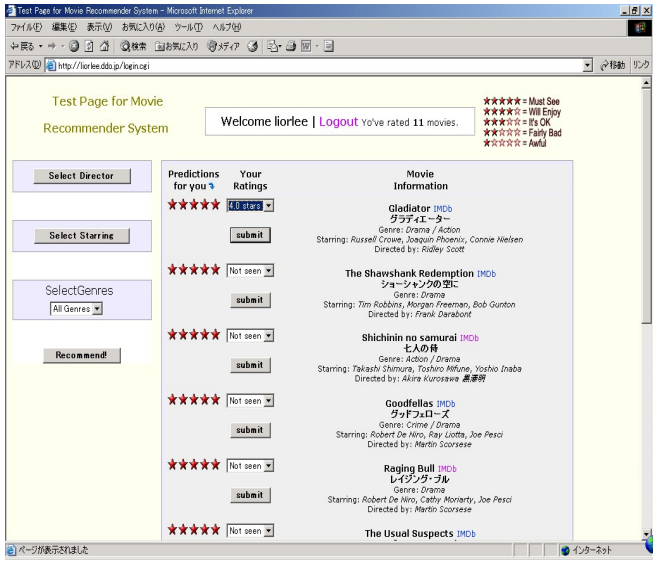

図 2 映画推薦システムのインタフェース

4.3 ユーザインタフェースと入力

図 2 はシステムのユーザインタフェースである. システムにログインするとまずこのページにたどる. ここで推薦されたアイテムは, ユーザプロファイルを基に C4.5 によって推薦された. 1 ページに 10 個のアイテムを表示し, Page2, Page3 と続く. “Predictions for you”はシステムの推薦度を表している. 星の数が多ければ多いほど推薦度が 高い. “Your Ratings”はユーザの評価で, 既に見ているアイテムがあれば, セレクトボックスで評価値を選択し, “Submit”ボタンでサーバに提出できる. “Movie Information”には映画のクレジット情報が表示されている. 英文タイトルの横のある“IMDb”をクリックすると Internet Movie Database(http://www.imdb.com)のそのタイトルのページに飛び, 詳しい情報を参照できる. これはシステムの有効性を考慮した設計である[7].

左側にある“Select Director”, “Select Starring”“Select Genres”はクレジットプレファレンスを与える時に使用する. “Recommend!”ボタンをクリックするとクレジットプレファレンスを用いた新しい推薦結果を表示させる.

一回ログアウトすれば, これまでに評価したアイテムが記録され, 次回の推薦のトレーニングデータとして用いられる.

4.4 C4.5 を用いた推薦リストの生成

C4.5 を使ってアイテムのコンテンツプレファレンス(7 つの属性, 5 段階の属性値)とそのアイテムに対するユーザの評価(5 段階の評価値)をトレーニングデータとして決定木を構築する. 映画データベースに格納されている未評価の映画は C4.5 のテストデータとして決定木によって分類され, 分類されたクラスに対応する予測評価(5 段階の評価値)の高い順に推薦候補リストが生成される.

4.5 クレジットプレファレンスによる推薦候補リストのソート

クレジットプレファレンスはジャンル, 監督, 出演者といった情報である. クレジットプレファレンスが推薦で用いられると, 推薦エンジンはまずジャンルが指定されているかどうかを調べ, 指定されていれば推薦候補リストの中から指定されたジャンルに属さないアイテムを取り除く. つぎに指定された監督および出演者が監督・出演したアイテムをすべて収集し, 各アイテムのコンテンツ属性 の平均値 を計算する. 推薦候補リストの中の同じ予測評価値を持ったアイテムに対して 値のもっとも高い属性 でソートを行う. そして の属性値が同じであるアイテムに対して 値が二番目に高い属性 でソートする. 例えば好きな監督に“Steven Spielberg”を選ぶと, 同じ 5 の予測評価値を持っているアイテムでも, ストーリ展開の面白さによってソートされ, 同じストーリ展開の面白さの属性値を持っているアイテムはさらに演技性によってソートされる.

コンテンツプレファレンスを使った推薦はトレーニングデータを増やさない限り, スタティックな結果しか得られないが, クレジットプレファレンスを指定することによってダイナミックな推薦結果をユーザに提示することが可能になった.

4.6 推薦結果の提示

推薦エンジンでは, 決定木によって生成された推薦候補リストを, クレジットプレファレンスを用いてソートし, その結果を, GUI を通してユーザに提示する. ユーザは提示されたアイテムに対して評価を行うことができ, 評価されたアイテムはトレーニングセットに加えられ, 次回の推薦に影響を与える.

5. 評価実験

5.1 実験目的

帰納学習を用いた推薦手法の有効性と推薦精度を検証するための評価実験を行う.

5.2 実験者と実験方法

実験者:大学生 20 人. 年齢層 19～23 才. 技術的背景:理工学系 14 人, 文系 6 人

データセット:象のロケット[8]から取得したコンテンツプレファレンスを持つ映画 1000 タイトル

実験方法:

① 適当なユーザ名でオンライン登録する
② 提示されたアイテムに対して評価を行う
③ 推薦リストのレビュー
④ リストに既に見ているアイテムがあれば, それについて評価を行う
⑤ クレジットプレファレンスを与え, ③, ④を繰り返す
⑥ 一定数のアイテムの評価を終えたら実験を終了する

5.3 評価基準

統計的評価基準は推薦予測値と実際のユーザの評価値を比較することによってシステムの推薦精度を評価する. 本論文では, もっとも広く使われている統計的評価基準としてのMAE(Mean Absolute Error)をシステムの評価基準とする. MAEの計算式は

$$MAE = \frac{\sum_{i=1}^{N} | p_i - q_i |}{N}$$

$< p_i, q_i >$は実際の評価値—予測値のペアで, N は評価されたアイテムの数である. MAE が低ければ低いほど, 推薦システムの精度が高いことを示す.

5.4 実験結果

図 3 にユーザ 20 人が行った映画推薦システムの評価実験の結果を示す. MAE の平均値は 0.63 で, 標準偏差は 0.37 である.

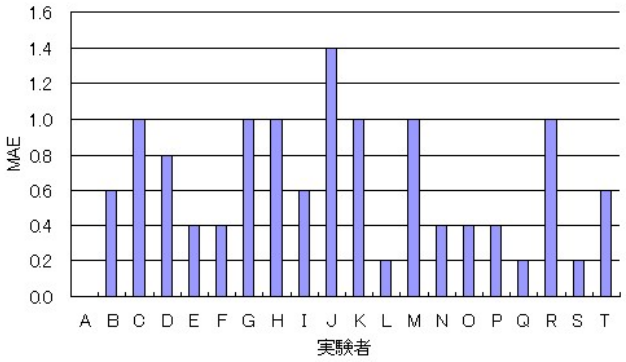

図 3 評価実験結果

5.5 実験の考察

本システムでは, 10 個という少ない評価項目数を用いてMAE を 0.7 以下に抑えることができた. [Sarwar,2001]らが行ったMovieLensの評価実験[3]では, MAEの値は0.7~0.8であった. データセットと評価方法が異なるので定量的な比較をすることはできないが, MAE が 1 以下であることは, ほとんどの場合, システムの予測値はユーザの実際の評価値と一致しているか, 1ランクずれているだけである. よって, システムはユーザの嗜好に適した推薦を行えたと言える.

また, オンラインで実験を行ったが, その際のサーバプログラムの実行速度は 2〜3 秒で, これはユーザに負担にならない速度だと言える.

問題点としては, 今回は大学生 20 人を実験者としたが, 年齢層が広がって, ユーザ数が増えると, 必ずしも同じ結果が出るとは限らないので, 実験結果の信頼性を高めるには実験データ数を増やす必要がある.

ユーザごとの評価値に格差があったので, MAE が高かったケースを調べた. トレーニングデータを与えるとき, 評価値に偏りがあると, 推薦精度が落ちることが分かった.

決定木による分類で, 分類に失敗したテストデータの割合は14%でした. より正確な分類を行うためには, アイテムのコンテンツプレファレンスの種類を増やす必要がある. 属性のストーリ展開の面白さが主成分であるとき, その評価値が低くても全体の評価値が高く付けられていることが多く見られた, 実験者の感想によれば, ストーリ性に欠けても, 印象的なシーンがいくつかあれば高く評価されることが分かった. よって, コンテンツプレファレンスの種類の増加において, 印象的なシーンという属性が考えられる. また, 本や音楽CDといったカテゴリの分類を行う際, 属性の数が少ないと, 判別空間に引っかからない場合があるので, より細かい分類が要求される. 今回の実験結果を参照すると, 分類に失敗したテストデータの割合が 10%以下になる分類の仕方が望ましい.

6. まとめ

本論文では, 推薦システムの要素技術としての協調フィルタリング手法の問題点を試みるに当たり, 帰納学習を用いた推薦手法を提案した. 帰納学習を用いた推薦手法では, ユーザ間で共有できるアイテムのコンテンツプレファレンスを推薦に用いることでスパーシティの問題を解決する. また, 学習時の計算コストは, 実際的に十分に小さいものであり, 瞬時にユーザプロファイルを生成できる. さらに, スケーラビリティの増大に伴って, 推薦時の計算コストが膨大になることもない.

帰納学習を用いた推薦手法の有効性を検証するために映画推薦システムを実装した. 実装したシステムをWWW(http://liorlee.ddo.jp)に公開し, オンラインで評価実験を行った.

帰納学習を用いた映画推薦システムでは, C4.5 による推薦に加えて, アイテムのクレジットプレファレンスを推薦候補リストのソートに用いることによって柔軟な推薦が可能になり, システムのトランスペレンシーの向上にも繋がる. そして, 数少ない評価項目を用いるだけで精度の高い推薦をユーザに提供できることが分かった.

協調フィルタリング手法を用いた推薦システムでは, ユーザ数と情報量のすさまじい増加に伴って, データの希薄性の問題による推薦精度の低下と, スケーラビリティの問題による推薦スループットの低下といった問題を抱えている. 本論文で提案した推薦手法はそれらの問題解決に取って有効であり, 帰納学習を用いた推薦システムはハイパフォーマンスな推薦をユーザに提供できることは評価実験によって確認した. 本論文は新しい推薦手法への試みであって, 意義のある研究と考えられる.

参考文献

[1] P. Resnick and H. R. Varian: ”Recommender Systems”, Communications of the ACM, vol.40, pp.56-58 , 1997

[2] M.Balabanovic and Y.Shoham : “ Fab : Content-based , collaborative recommendation. ”, Communications of the ACM, vol.40, no.3, pp.66-72, 1997

[3] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J.: “ Item-Based Collaborative Filtering Recommendation Algorithms”, In Proc. of the 10th International World Wide Web Conference (WWW10), Hong Kong, 2001

[4] 豊田 秀樹:“金鉱を掘り当てる統計学—データマイニング入門”, ブルーバックス, 2001

[5] Ross J. Quinlan. C4.5. “Programs for Machine Learning.”, Morgan kaufmann Publishers Inc.,San Francisco, California, 1993

[6] N. Abe and H. Mamitsuka : “Query Learning Strategies Using Boosting and Bagging”, Proc.of the 15th Int. Conf. on Machine Learning (ICML98), pp:1-9, 1998

[7] Kirstn Swearingen, Rashmi Sinha: “Beyond Algorithms: An HCI Perspective on Recommender Systems.”, ACM SIGIR Workshop on Recommender Systems, 2001

[8] www.paoon.com