協調フィルタリングを用いた 映画評価予測プログラムの試作 -ジャンル嗜好による協調フィルタリング-

Trial program for movie rating prediction using Collaborative Filtering -Collaborative Filtering by genre taste-

谷　研究室　　吉野　友美
Yoshino Tomomi

概要

協調フィルタリングを用いて、ジャンル分けされた映画のタイトルによるレコメンドシステムを試作する。

1　はじめに

オンラインショッピング等では、ユーザーの購買意欲を向上させるためにお薦め商品の提案がなされている。そしてユーザへその提案をするためにレコメンドシステムが使われている。今回は協調フィルタリングを用いて、映画評価予測プログラムの試作を行った。

今回はジャンル分けされたデータを使って、ユーザがどのジャンルを好むかという嗜好を予想して、それを元に評価することに挑戦する。

２章では協調フィルタリングの定義をして、３章ではアルゴリズムの方針を提示し、４章で実験をして、５章ではその結果、６章で今後の課題を示していく。

2　協調フィルタリング

協調フィルタリング（Collaborative Filtering, CF）は、多くのユーザの嗜好情報を蓄積し、あるユーザと嗜好の類似した他のユーザの情報を用いて自動的に推論を行う方法論である。趣味の似た人からの意見を参考にするという口コミの原理に例えられることが多い。

例えば、ユーザ A がアイテム X を好むとすると、アイテム X を好む別のユーザ B が好むアイテム Y を探し出し、ユーザ A もアイテム Y を好むのではないか、という推論をコンピュータによって自動的に行う。実装にはユーザ同士の類似度を、同じアイテムにつけた評価の相関係数などによって表して類推に利用することが多い。

2.1　ユーザーベース

ユーザーベースとは、推薦対象のユーザーと他のユーザー間の類似度を求め、推薦対象のユーザーの未評価値を他のユーザーとの類似度とそのユーザーの評価値を用いて予測するものである。

2.2　アイテムベース

アイテムベースとは、各アイテム間の類似度を求め、推薦対象のユーザーの未評価値を他のアイテムとの類似度とそのアイテムの評価値を用いて予測するものである。

3　方針

3.1　基本的な手順

まず基本的な手順を提示して、これに基づいてさらに改造を試みる。

始めに各ユーザーのジャンル毎の平均を求め、そのジャンルが好むか好まないか判断する。また映画のタイトルの属するジャンルからそのタイトルが好むか好まないか予想し、そのタイトルを好むであろう人の評価の集合の平均とそのタイトルを好まないであろう人の評価の集合の平均を求める。最後に未評価のタイトルに対しても同様に、そのタイトルが属するジャンルから好むか好まないか予想し、そのタイトルを好むであろうならば好む人の集合の評価の平均を割り当て、好まないであろうならば好まない人の集合の評価の平均を割り当てていく。

以下より手順の詳細を記述する。

(1) 各ユーザーが既に評価した映画のタイトルの評価をそのタイトルが属するジャンルに足していく。尚、属するジャンルが複数ある場合はそれぞれのジャンルに足していく。そして、各々のジャンルの評価の和から平均を求める。

(2) ジャンルの平均が 3.5 以上ならばそのジャンルを好み、3.5 未満ならばそのジャンルは好まないと判断する。

(3)(2) で作成されたジャンルの好みのデータを元に、映画のタイトルが属するジャンルが好むかどうかによって、そのタイトルを好むか好まないか判断する。
(4) 映画のタイトルが好むか好まないか判断した上で、好む人の評価の集合と好まない人の評価の集合に分けて、平均を求める。
(5) 未評価のタイトルに対して、そのタイトルが属するジャンルを好むか好まないかによって、そのタイトルを好むか好まないか予想して、好むならばそのタイトルを好む人の評価の平均を、好まないならばそのタイトルを好まない人の評価の平均を割り当てる。

3.2 さらに改造した点

基本手順に則り、少々改造を試みる。
(i) 上記で出力された結果と各ユーザーのジャンル毎の平均の平均を取る。
(ii) 好むと好まないの他に普通という基準を設ける。
１＜＝好まない＜３
３＜＝普通＜４
４＜＝好む＜＝５

4 実験

4.1 実験環境

CPU	Athlon64 3200+
メモリ	1024MB
OS	CentOS 5.2

4.2 データセット

アルゴリズムの性能評価に MovieLens のデータセットを利用する。
MovieLens のデータセットはユーザー数:943、映画数:1682、評価数:10 万で構成されていて、各ユーザーは評価数が 20 以上であることが保証されている。今回はこのファイルの中から各ユーザー毎に 10 個ずつ評価をランダムに抜き出し、予測対象データと元データに分ける。
ここで、コールドスタート問題を避けるため、各ユーザーの最低評価数 10 は保証することとする。
そのファイルを 10 組作り出した。そしてその 10 組のファイルを用いて実験を行い、精度を求め、その平均を比較する。

4.3 方法

精度を比較するための指標として MAE を用いる。

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | \text{予測値} - \text{実際の値} |$$

ここで、N は予測対象データの総数である (=9430)。MAE の値が 0 に近づけば近づくほど誤差が少ない、つまり精度が良いということになる。

5 結果・考察

5.1 結果

方法	精度
基本手順	0.861948
改造１	0.860496
改造２	1.200622

5.2 考察

基本手順ではそれなりの結果が得られたと思う。改造１では精度が上がったが、改造２では下がっていた。改造１で行った基本手順での結果にさらにジャンル平均との平均を取ったもので精度が上がったのは、ユーザーのジャンル平均も考慮に入れることで、個人個人の特徴が反映されたからだと思われる。逆に改造２で行った好む好まないの他に普通という基準を設けたもので精度が落ちたのは、タイプを分けすぎたことにより比べるユーザーの数が減り、平均にばらつきが見られたことが原因だと思われる。

6 今後の課題

今後、以下の課題が考えられる。
・Netflix のデータを扱うために、データベースの構築
・実行速度向上のために、マシンスペックの改善

7 参考文献等

[**1**] MovieLens Data Sets http://www.grouplens.org/node/73

[**2**] Netflix prize http://www.netflixprize.com/