九州工業大学 情報工学部 システム創成情報工学科 廣瀬研究室

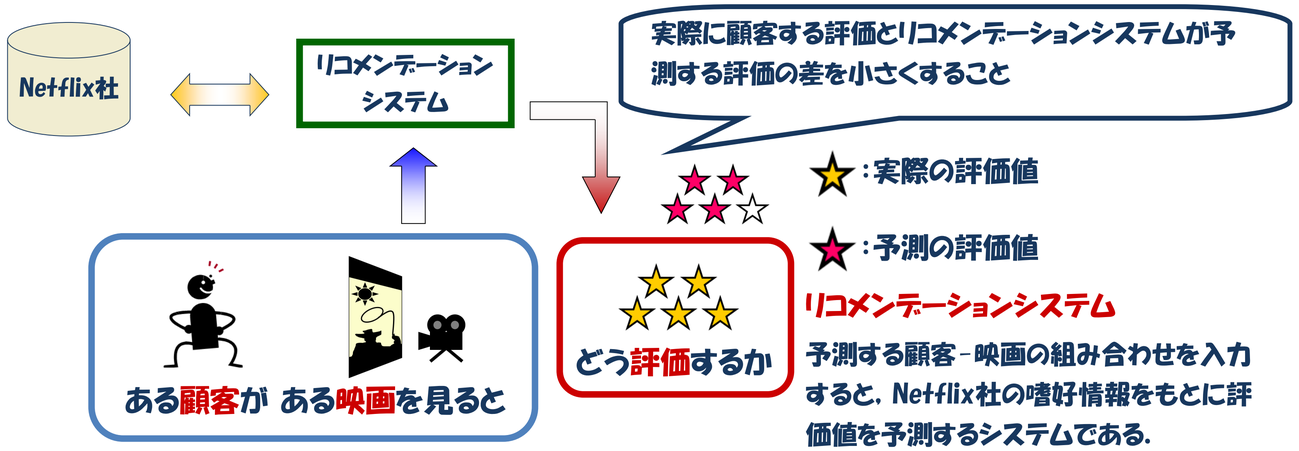
リコメンデーションアルゴリズムの研究

研究対象

オンラインDVDレンタルサービス会社のNetflix社は，膨大な情報の中からユーザの嗜好に合った情報を見つけ出し，その情報をユーザに能動的に提供する推薦技術の研究を行っている．また，自社の推薦システムの予測精度を高めるために，自社の顧客の嗜好データを公開し，多くの研究者と協調しながら研究を進めている．提供されたデータを研究対象とする．

研究目的

Netflix社に蓄積された顧客の嗜好情報をもとに，ある顧客がある映画を見るとどう評価するのかを予測する．この予測の**精度を高めること**が研究目的である．



予測手法

データ

Netflixデータは**478,615人**の顧客の**17,770本**の映画に対する**5段階評価**データである．評価されている項目は約**1億件**である．このデータを分析し，特徴を作成する．

マトリックス分解を用いた推薦アルゴリズム

映画ー顧客の評価を説明する**潜在的な特徴**を各の映画と各の顧客に置くことで予測を可能とする．潜在的な特徴は映画のジャンルかもしれません．特徴がジャンルとするなら，映画に置いた特徴はその映画のジャンルを表すことになる．そして，顧客に置いた特徴はその顧客のジャンルへの好みを表すことになる．このような潜在的な特徴を設定することはマトリックス分解技術を使うことで可能となる．マトリックス分解技術の1つである**特異値分解**を用いて特徴を抽出する．

$$\begin{bmatrix} Original_{I\times J} \end{bmatrix} \approx \begin{bmatrix} Customer_{I\times K} \end{bmatrix} \begin{bmatrix} Movie_{K\times J} \end{bmatrix} = \begin{bmatrix} \Pr edict_{I\times J} \end{bmatrix}$$

K : Number Of Latent Factor

潜在的な特徴の推定方法は，全ての既知評価に対して右に示す1～3の手順を行います．そして，全ての既知評価値について行えた場合に，操作手順4により二乗誤差が最小になっているかを判断する．最小になっていなければ同様の処理を行う．なっていれば次の特徴量に対しても同様の処理を行う．

①予測

$$\hat{x}_{ij} = \sum_{k=1}^{K} c_{ik} m_{kj}$$

②既知と予測の誤差

$$e_{ij} = x_{ui} - \hat{x}_{ij}$$

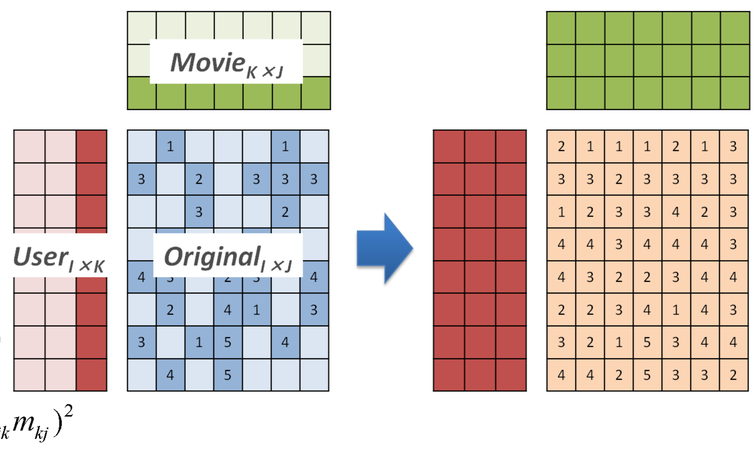
③c_{ik}とm_{kj}を更新

$$c_{ik} \leftarrow c_{ik} + \eta(e_{ij} m_{ki})$$

$$m_{kj} \leftarrow m_{kj} + \eta(e_{ij} c_{ik})$$

④条件の判定

$$\arg\min \sum_{i,j} (x_{ij} - \sum_{k} c_{ik} m_{kj})^2$$



推薦アルゴリズムの予測精度の評価

推薦アルゴリズムの予測精度の評価は予測の精度をあらわす指標*RMSE*を用いて行う．

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{X}_n)^2}$$

X : 既知評価
$\hat{X}$: 予測評価
N : サンプル数

特異値分解にリッジ回帰を適用

目的関数に正則化項を加えてることで，予測精度をさらに向上できる．

$$E = \frac{1}{2}\sum_{i,j}(x_{ij} - \sum_{k} c_{ik} m_{kj})^2 + \underbrace{\frac{1}{2}\lambda(\sum_{k} c_{ik}{}^2 + \sum_{k} m_{kj}{}^2)}_{\text{正則化項}}$$

SVDとリッジ回帰を適用させた場合の特徴量数とRMSEの関係図：