# 鑑賞履歴に基づいた映画推薦アルゴリズム

Recommendation Algorithms Based on Behavioral History.

王剣威
Wang Jianwei
指導教員　平原誠

法政大学大学院理工学研究科応用情報工学専攻修士課程

With the rapid popularity of Internet and mobile Internet, the number of movie entertainment information on the Web is quite huge, and it is increasingly difficult for people to obtain information about movies of interest. In this study, we propose a personalized recommendation strategy based on the sequence of user playing behavior for the personalized recommendation problem of movie websites. The strategy analyzes the user playing video behavior data by Word2vec, a deep neural network word vector model, maps the videos into equal-dimensional feature vectors, calculates the similarity of movies, and uses it as the basis of recommendation to generate a recommendation list to users. In addition, to improve recommendation accuracy, Word2vec parameter settings were considered, and a pre-processing method for history sequences was proposed.

***Key Words:*** *recommendation, Word2Vec*

## １．はじめに

自然言語処理における分散表現技術である Word2vec[1]は、テキスト文書から単語ベクトルを学習する方法である。自然言語処理タスクで良好な結果を得ており、近年では商品推薦システム[2]にも利用されるようになってきた。Word2vec の手法を用いた推薦システムでは、学習で得られたベクトル間の類似度を計算することで、推薦処理を実現する。Oren Barkan[3]は音楽鑑賞と Microsoft App Store の商品履歴に Word2vec を適用した。しかしながら、Word2vec のパラメータをどのように設定すべきか、Word2vec の精度をどのように向上させるかについては検討されていない。本研究では、映画鑑賞の履歴シーケンスから有向グラフ（履歴グラフ）を自動生成する手法を提案し、その履歴グラフをランダムサンプリングすることで生成される擬似履歴シーケンスを学習に用いることで、推薦精度を向上させることを目的とする。加えて Word2Vec のパラメータであるベクトル表現の次元数の変化が推薦精度にどのような影響を与えるかを検討する。

## ２．Skip-gram

Skip-gram は図 1 に示すように、入力単語$\boldsymbol{w_t}$（中心単語）からその周辺単語$\{w_{t-i},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+i}\}$を予測するニューラルネットワークである。以下では、中間層の次元を N とする。

単語ベクトルは学習後の Skip-gram における中間層への重み行列から取得できる。入力ベクトルは one-hot 表現であるため、ある単語ベクトルは重み行列の当該単語の行成分となる。

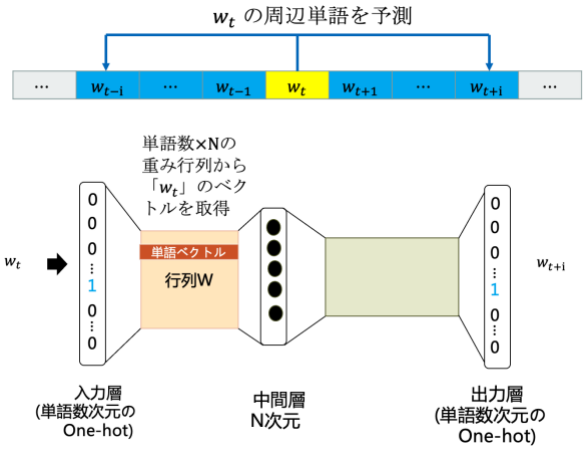


図 1 Skip-gram のニューラルネットワーク

## ３．提案手法

本研究では、ユーザーの映画の鑑賞履歴を自然言語処理における単語系列として扱い、Word2vec を用いて学習する。これにより個々の映画の特徴が反映された映画ベクトル生成し、それらに基づいて好みの映画を推薦するアルゴリズムを提案する。

### ３．１　実験データ

本研究では, movie lens 研究チームが企画した movielens20M データセットを使用する。利用者 13 万 8000 人による 2 万 7000 本の映画に対するスコア 2000

万件が含まれている。データセット内には映画の ID とその名称および５段階レーティング(0、0.5、1、…、4.5、5)が含まれている。図 2 にレーティングの頻度分布示す。

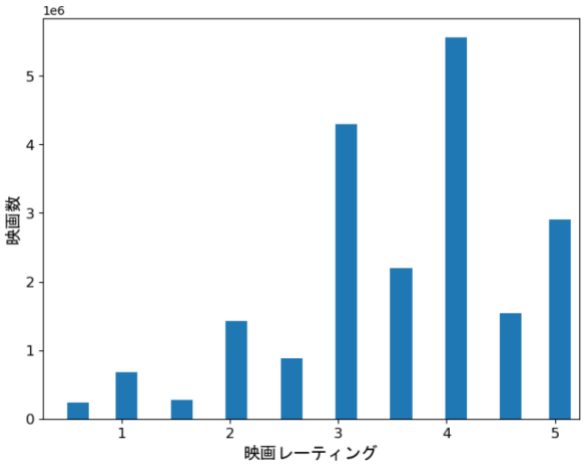


図 2 映画レーティングの頻度分布図

### ３．２　前処理

映画ベクトルを学習する前に、Word2vec における「単語」と「文章」に相当するデータをデータセットから取得する必要がある。本研究では、それぞれの「映画」を「単語」とみなし、ユーザーが鑑賞した映画のシーケンスを「文章」とみなす。具体的には、レーティング 4 以上の全ての作品をユーザーの「好きな映画」とし、「好きな映画」のみを格納したシーケンス（履歴シーケンス）を作成して、「文章」とみなす。

### ３．３　履歴グラフ生成アルゴリズム

ユーザーの履歴シーケンスが短いと、映画ベクトルの質が低下する恐れがある。本研究では図３に示すように全ユーザの履歴シーケンスに基づいて、個々の映画を表すノードと映画の鑑賞順を反映した有向エッジとからなる有向グラフ（履歴グラフ）を自動生成する。そして、この履歴グラフを利用し、ランダムに選んだ映画ノードを始点としてランダムサンプリングすることで、擬似履歴シーケンスを生成し、Skip-gram で学習を行う。

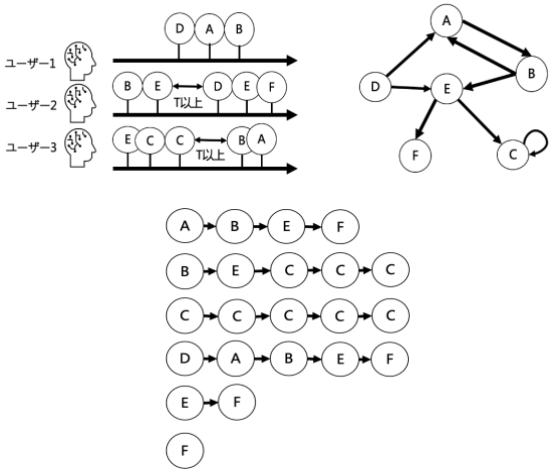


図３　擬似履歴シーケンス

### ３．４　結果評価

各ユーザーの履歴シーケンスを訓練シーケンスとテストシーケンスに 7：3 の割合で分割した。訓練シーケンスを用いて映画ベクトルを生成し、映画間の類似度を計算する。映画$i(=1,\cdots,L)$の m 次元映画ベクトルを$v_i^x(x=1,\cdots,m)$で表すと、映画$i$と$j$との類似度は

$$W_{ij}=\frac{\sum_{x=1}^{m}\left(v_i^x\times v_j^x\right)}{\sqrt{\sum_{x=1}^{m}\left(v_i^x\right)^2}\times\sqrt{\sum_{x=1}^{m}\left(v_j^x\right)^2}} \quad (1)$$

と表すことができる。各ユーザーの訓練シーケンスに含まれる映画と類似度が高い上位 N 本の映画（Top-N）を当該ユーザーへの推薦リストとする。それぞれのユーザに対して推薦リストとテストシーケンスに含まれる映画を比較することで適合率、再現率、F1 スコアを用いてモデルの性能を評価する。

## ４．結果と考察

本研究では、映画ベクトルの次元の違いが推薦の効果に与える影響を実験的に検証した。推薦リスト Top-N=20 とした場合の適合率、再現率、F1 値を図４に示す。

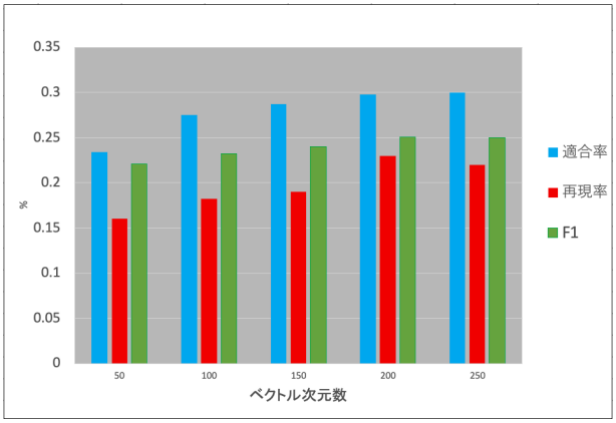


図４　映画の分散表現の次元が推薦精度に与える影響

図 4 に示すように、映画ベクトルの次元が大きくなるにつれて推薦の適合率が向上し、200 以上の次元になると横ばいになっていることがわかる。再現率と F1 値も同様の傾向にあることが分かる。映画ベクトルの次元が小さすぎると、映画ベクトルが映画の本質的な情報を十分に反映できないことは容易に理解できる。ある程度まで次元を上げると、映画ベクトルが映画情報を完全に表現してしまい、推薦精度の大幅な向上は見込めなくなる。また、次元数の増加に伴い、学習や推薦リスト生成に時間がかかるため、計算速度と推薦精度を天秤にかけて次元数を選択する必要がある。後続の実験では、映画ベクトルの次元を 200 とした。

本研究では、従来手法（item2vec）と提案手法との比較実験を行った。 推薦リスト Top-N={5,10,15,20,25,30}を独立変数として求めた適合率、再現率、F1 値を図５に示す。

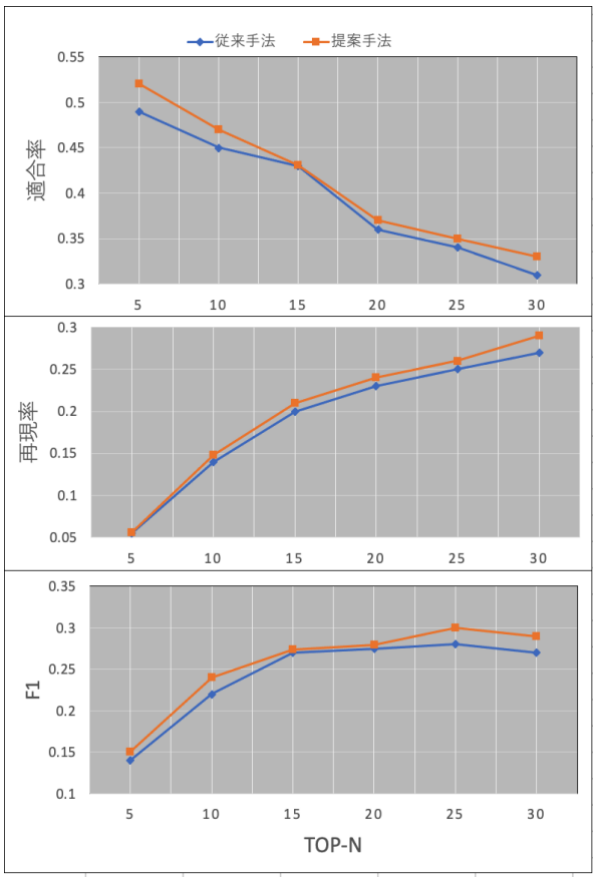


図５ Top-N の違いが推薦精度に与える影響

図５から、(1)Top-N が大きくなると、ユーザーに推薦する映画リストに含まれる映画の数が増えるため、再現率が上がり、適合率が下がることがわかった。 (2)F1 指標は Top-N とともに上昇し、Top-N=25 で最高値に達して、その後下降していることがわかる。 (3)提案手法は従来手法に比べて推薦精度が若干良いが、大きな差はないことがわかった。

## ５． 今後の展望

提案手法は、映画鑑賞履歴から自動生成される履歴グラフをランダムサンプリングすることにより擬似履歴シーケンスを作り出し、それらを Skip-gram の学習に用いるという考えに基づいている。 今後の研究では、履歴グラフのノード間に重みを導入し、精度を向上させる必要がある。これに加えて、ユーザーの行動や性格の類似度を考慮する必要があるだろう。例えば、映画を鑑賞する際、映画を早送りして楽しむユーザーと、通常の速度で楽しむユーザーとでは、映画に対する好みの程度が一致しない。したがって、将来的にはユーザーの行動や性格の類似度を考慮した映画推薦アルゴリズムの研究を進めていく。

**謝辞**

本研究を進めるにあたり，終始適切な助言を賜り，また丁寧に指導してくださった，修士論文指導教員の平原誠准教授に心より感謝いたします.

**参考文献**

1) Church, Kenneth Ward. "Word2Vec." Natural Language Engineering 23.1 (2017): 155-162.
2) 神嶌敏弘：推薦システムのアルゴリズム,人工知能学会誌, Vol. 23, pp. 248–263, 2008.
3) Barkan, Oren, and Noam Koenigstein. "Item2vec: neural item embedding for collaborative filtering." 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2016.
4) 森純一郎：ウェブ情報を用いたエンティティのランキング学習に関する研究,人工知能学会全国大会論文集, 第 24 回全国大会,2010.
5) SIG-KDD, A. C. M. "DeepWalk: Online Learning of Social Representations." (2014).