# デジタルアーカイブのためのグラフベースの深層学習による推薦システム

王 嘉韻（立命館大学 情報理工学研究科）
Biligsaikhan Batjargal（立命館大学 衣笠総合研究機構）
前田 亮・川越 恭二（立命館大学 情報理工学部）
赤間 亮（立命館大学 文学部）

本論文では，人文系デジタルアーカイブによく見られるグラフ構造のデータに適したグラフベースの深層学習を用いた推薦システムを提案する．提案手法の有効性を検証するため，グラフベースの深層学習推薦アルゴリズムを利用し，立命館大学アート・リサーチセンターの浮世絵データベース（以下 ARC-UDB という）のデータで検証した．その結果，従来の推薦手法と比較して予測精度が大幅に改善された．本研究の新規性は以下である：1）グラフベースの深層学習推薦アルゴリズムをデジタルアーカイブに応用した．2）深層学習モデルの入力ベクトルを改良し，浮世絵推薦のタスクにより適したものにした．3）比較実験を行い，提案手法の浮世絵データに対する有効性を証明した．この提案手法は，浮世絵だけでなく，他のグラフ構造のデータセットへの使用も考えられる．

# Graph-Based Recommender System Using Deep Learning for Digital Archives

Jiayun Wang (Graduate School of Information Science and Engineering, Ritsumeikan University)
Biligsaikhan Batjargal (Kinugasa Research Organization, Ritsumeikan University)
Akira Maeda / Kyoji Kawagoe (College of Information Science and Engineering, Ritsumeikan University)
Ryo Akama (College of Letters, Ritsumeikan University)

In this paper, we propose a recommender system using a graph-based deep learning method that is suitable for graph-structured datasets which are often found in humanities digital archives. In order to verify the effectiveness of the proposed method, we used a graph-based deep learning recommendation algorithm and verified it with data from the ukiyo-e database of the Ritsumeikan University Art Research Center (ARC-UDB). As a result, the prediction accuracy was significantly improved compared to the conventional recommendation method. The novelty of this study is as follows: 1) A graph-based deep learning recommendation algorithm was applied to a digital archive. 2) The input vector of the deep learning model was improved to make it more suitable for the ukiyo-e recommendation task. 3) A comparative experiment was conducted to prove the effectiveness of the proposed method for the ukiyo-e dataset. The proposed method can be used not only for ukiyo-e, but also for other datasets with graph-like structures.

## １．まえがき

ほとんどのデジタルアーカイブと同様に，Ritsumeikan University Art Research Center（ARC）浮世絵データベース[1]には検索機能が実装されている．しかし，ユーザがデジタルアーカイブを閲覧する際，検索機能を使用する以外にも，興味のある推薦情報（たとえば閲覧履歴に基づいた推薦結果，閲覧中の内容に基づいた推薦結果など）を取得したい場合がよくある．本論文では，増加するデジタルアーカイブの訪問者の情報ニーズに応え，ユーザがデジタルアーカイブに保存された情報を十分に活用できること，またデジタルアーカイブを使用する時により多くの情報を取得できることを目指し，推薦機能を新たな情報獲得手段として提案する．

浮世絵は，17 世紀から 19 世紀にかけて普及した日本の芸術の一つであり，ARC 浮世絵ポータルデータベース（以下 ARC-UDB という）では 162,521 枚（2019 年 10 月現在）の浮世絵とその情報を公開している．

メタデータは常にデジタルアーカイブを利用するユーザにとって重要な要素であるため，本研究では主にメタデータに基づいた推薦に焦点を当てる．本研究で扱うデータは，ARC-UDB のロ

 

| f1 | f10 | f11 | f12 | f13 | f14 | f15 | f2 | f20 | f21 | f22 | f23 | f27 | f28 | f29 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| arcUP2977 | | 1 | 伊場仙板 | 773 | 111 | 伊場屋　仙三郎 | arcUP2976 | 大判/錦絵 | 横 | 一勇斎国芳画( | 国芳 | 1843 | 天保１４ | ~弘化０４ |
| arcUP4368 | arcUP4368 | 1 | | | | | arcUP4368 | 大判/錦絵 | 横 | － | 国周 | | | |
| arcUP2435 | | 1 | 錦鯱堂 | 上84 | 5084 | | arcUP2435 | 大判/錦絵(杉 | 横 | 広貞 | 広貞 | 1851 | 嘉永０４ | |
| arcUP2451 | | 1 | － | － | | | arcUP2451 | 中判/錦絵 | 横 | 広貞 | 広貞 | 1849 | 嘉永０２ | |
| arcUP2452 | | 1 | － | － | | | arcUP2451 | 中判/錦絵 | 横 | 広貞 | 広貞 | 1849 | 嘉永０２ | |
| arcUP2444 | | 1 | 断裁 | | | | arcUP2444 | 中判/錦絵 | 横 | 広貞(印) | 広貞 | 1849 | 嘉永０２ | |
| arcUP2445 | | 1 | － | | | | arcUP2444 | 中判/錦絵 | 横 | 広貞 | 広貞 | 1849 | 嘉永０２ | |
| arcUP2493 | | 1 | 力[上田] | 上71a | 5071 | | arcUP2493 | 中判/錦絵 | 横 | 広貞 | 広貞 | 1849 | 嘉永０２ | 頃 |
| arcUP2494 | | 1 | 力・上田 | 上71a | 5071 | | arcUP2493 | 中判/錦絵 | 横 | 広貞 | 広貞 | 1849 | 嘉永０２ | 頃 |

図 2　浮世絵メタデータの一部
Figure 2　A part of ukiyo-e metadata.

グデータ（閲覧記録，図 1）と浮世絵のメタデータ（浮世絵を説明するデータ，絵師や画題など，図 2）である．データの詳細とデータ処理の流れは第 4 章で述べる．

```
121.105.110.62 - - [08/Apr/2018:04:04:25 +0900] "GET /favicon.ico HTTP/1.1" 404 209 "http://www.dh-jac.net/db1/books/results-detail.php?f1=SOA41983&f39[]=00%2F000&f65[]=%E5%A4%A9%E8%A1%A3%E7%B4%9B%E4%B8%8A%E9%87%8E%E5%88%9D%E8%8A%B1&-max=15&singleskip=0&enter=shochikudaihon" "Mozilla/5.0 (Linux; Android 8.0.0; HTV33 Build/OPR6.170623.013; wv) AppleWebKit/537.36 (KHTML, like Gecko) Version/4.0 Chrome/65.0.3325.109 Mobile Safari/537.36"
```

図 1　ARC-UDB のログデータの一例
Figure 1　An example of ARC-UDB log data.

一枚の浮世絵は複数のメタデータ項目を持ち，また，そのメタデータは複数の浮世絵に共通である場合がある．例えば，浮世絵作品「神奈川沖浪裏」のメタデータ項目の一つは絵師であり，その値は「葛飾北斎」である．同時に，「葛飾北斎」が絵師である浮世絵（「隅田川関屋の里」など）は複数存在する．このような many-to-many 関係は通常グラフ構造で単純かつ明確に表現することができる（図 3）．同様に，ユーザと浮世絵の関係も many-to-many である（図 3 に示すように，

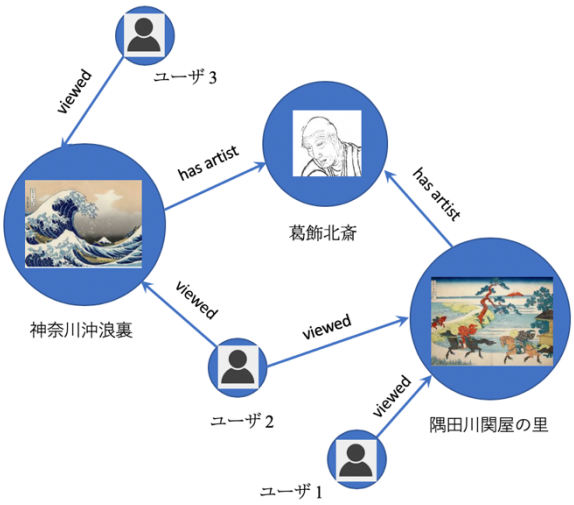


図 3　ARC-UDB における many-to-many 関係の例
Figure 3　An example of many-to-many relationships in ARC-UDB.

一人のユーザは複数の浮世絵を閲覧し，一枚の浮世絵は複数のユーザに閲覧される）ため，グラフベースの推薦アルゴリズムは ARC-UDB（他の同様なデジタルアーカイブも含む）における推薦に適していると考えられる．

## 2．関連研究

基本的なグラフベースのアルゴリズムを利用する推薦システムは，数十年にわたって研究されてきた[2，3]．グラフは，ノードとエッジによってアイテム間，ユーザ間，ユーザとアイテムの相互関係を表すことができる．したがって，単純な協調フィルタリングなど，その相互関係だけを利用する伝統的な推薦方法はグラフ構造のデータに容易に適用できる．

また，グラフでは，構造化データ（メタデータなど）を容易にノードに埋め込むことができる．Musto らの研究[4]では，推薦のために構築したグラフ構造に，Web から抽出した linked open data（LOD）を埋め込み，比較実験を行った．その結果，追加した LOD データは適切にアイテムを表現することができ，グラフにより新たな知識を獲得し，従来手法と比較して高い推薦精度を得られることを示した．

## 3．グラフベース推薦システム

前述の通り，推薦システムでは，ユーザとアイテムの関係は many-to-many 関係である．推薦システムでは，ユーザとアイテムの関係は基本的であり，メタデータなどユーザとアイテムを表す情報はユーザノードとアイテムノードに埋め込むことができるため，図 3 をさらに抽象的かつ簡潔に表すと図 4 のようなグラフになる．

図 4 では，ユーザノード，アイテムノードとその関係（閲覧）をそれらの間のエッジで表している．実線は既存の関係を意味し，点線は潜在的な関係を意味する．

グラフ内のノード間の潜在的な関係を予測する手法として，リンク予測がある．例えば図 1 では，ユーザ B はアイテム A のみを閲覧しており，ユーザ B がアイテム B やアイテム C に興味があるかどうかは分からない．リンク予測は，ユーザ

B とアイテム B，およびユーザ B とアイテム C の近接度を分析することにより，その間にリンクがあるかどうか（ユーザがあるアイテムに興味があるかどうか）を予測する．

ユーザとアイテムの間のリンクには属性がある．ここでは，ユーザはアイテムを閲覧したという意味で，属性を「閲覧」にしている．実際にグラフ関係をアルゴリズムで扱う際には，ユーザ対アイテムの嗜好度を表すことができる数値を利用する．詳細な内容は第 5.1 章で述べる．

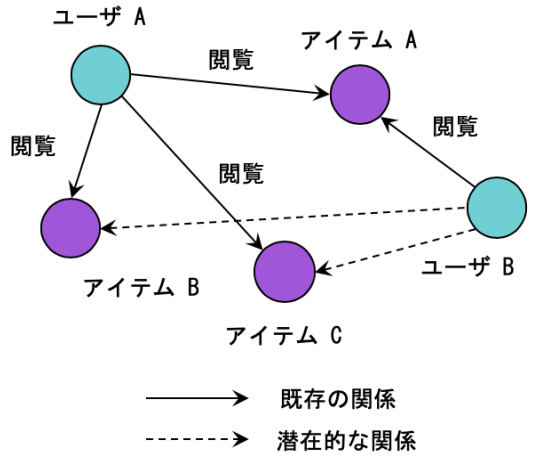


図 4　グラフによるユーザとアイテムの関係
Figure 4　Relationship between users and items by a graph

## 4．データセットとその処理

本研究で扱う浮世絵データとそれぞれのデータセットの詳細および，それらに対する前処理を以下に述べる．

**ログデータ（閲覧記録）**：本論文で使用するログデータは 2014 年 8 月から 2018 年 4 月までの有効ログデータ約 103 万件であり，このうち本研究で不要なデータをフィルタリングした結果 11.5 万件が残った．このデータから各々のユーザの閲覧状況が分かる．ARC-UDB には評価システムが実装されていないため，本研究では，ユーザがある浮世絵を閲覧した回数により，どの程度ユーザがその浮世絵を気に入ったかを評価する．閲覧回数が多いほど，ユーザがその浮世絵をより好んでいると仮定する．ログデータはユーザの識別，ユーザ対アイテムの評価，閲覧した Web サイトのスクレイピングのために用いる．図 1 に示したログデータの一例のフォマットは「IP – time stamp – requirement – status – web page size – source URL – agent name」となる（図 5）．

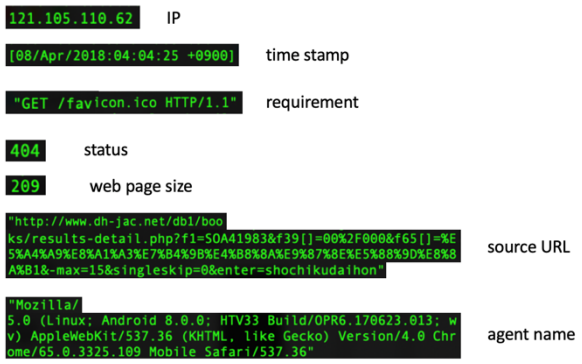


図 5　ログデータの解釈
Figure 5　Interpreting access log data

ここで，「IP」は IP アドレス，「time stamp」はあるユーザがある Web ページにアクセスした日時，「requirement」はユーザが要求したコンテンツである．たとえば，ユーザが特定の Web ページへのアクセスを目的としている場合，requirement はその Web ページの URL になる．

「status」はアクセスが成功したかどうかを示す数値（HTTP ステータスコード）である．「200」は，正常にアクセスできたことを意味する．

「web page size」は，アクセスされた Web ページのサイズを示す数値である．

「source URL」はどの Web ページから今の要求（requirement）が行われたかを示す．

「agent name」はアクセスしたユーザのブラウザ名とバージョンなどを示す．Agent name を用いてユーザがロボットかどうかを判断することができる．ログデータの処理手順は以下の通りである：

1. Status「200」の成功したログレコードのみを保持する．
2. 「bot」，「crawler」，「spider」など，agent name に存在するいくつかの特別な単語によってロボットユーザを削除する．ARC-UDB サーバではロボットユーザとして合計 102 の単語を登録している．
3. 1 つの浮世絵の詳細情報を示す Web ページにアクセスする requirement のみを保持する．これらの requirement を選択する理由は，ARC-UDB では，検索後に単一の浮世絵画像をクリックした場合にのみ，特定の浮世絵の詳細が表示され，これがあるユーザが特定の浮世絵に興味があることを示すと考えられるためである．
4. ARC の編集者からのアクセスログを削除する．編集者のみが ARC-UDB の編集モードを使用できるため，パターン「XXX / edit

/ XXX」を含むログレコードを単純に削除する．ただし，編集者は編集モードなしで Web ページをチェックする場合がある．編集者が一般ユーザとして表示されることを避けるため，訪問数が限られているユーザのログデータのみを使用する．具体的には，次節で詳しく説明する．

**浮世絵のメタデータ**：本研究において，メタデータとは，浮世絵の絵師，分類，画題など，ARC-UDB の浮世絵の属性を説明するデータを意味する．本論文では，メタデータ項目「artist」（絵師）を用いて初期のユーザベクトルとアイテムベクトルを構築する．

## ５． 提案手法

### ５.１フレームワーク

図 6 に提案手法の流れを示す．点線内は提案する推薦システムである．ユーザの行動はユーザインタフェースによってログデータに保存する．ARC-UDB の浮世絵データセットには画像とメタデータが保存されている．本論文では予めメタデータのみを考慮して推薦アルゴリズムを開発する．浮世絵推薦システムはログデータとメタデータの必要な部分を用いてトレーニングデータとして扱う．そして，トレーニングデータを用いて HinSage モデルを訓練し，リンク予測の結果を獲得する．その結果は各々のユーザの各々のアイテムに対する嗜好度である．最後に，推薦アイテム選択部が予測した嗜好度の高いアイテムを推薦内容としてユーザに提示する．

トレーニングデータは ARC-UDB のログデータから得られたグラフのノードペアとそれに対応するユーザ対アイテムの嗜好度である．嗜好度の計算は，ログデータから得られたユーザがアイテムを閲覧する頻度を標準化した値である．標準化した値は以下の式で求める：

$$nvf = \log_2(ViewFrequency) + 1$$

「ViewFrequency」は元の閲覧頻度である．「nvf」は normalized view frequency, 標準化した閲覧頻度を意味する．

トレーニングデータ中で，インプット特徴は結合したノードペアのユーザ特徴（ベクトル）とアイテム特徴（ベクトル）であり，正解データ（ラベル）は nvf 値である．例えば，ユーザ 1 がアイテム 1 を 4 回閲覧した場合，nvf 値は 3 になる．インプット特徴はユーザ 1 の特徴とアイテム 1 の特徴の結合になる．

ユーザ特徴とアイテム特徴の埋め込みは次節で述べるこれから記述する．

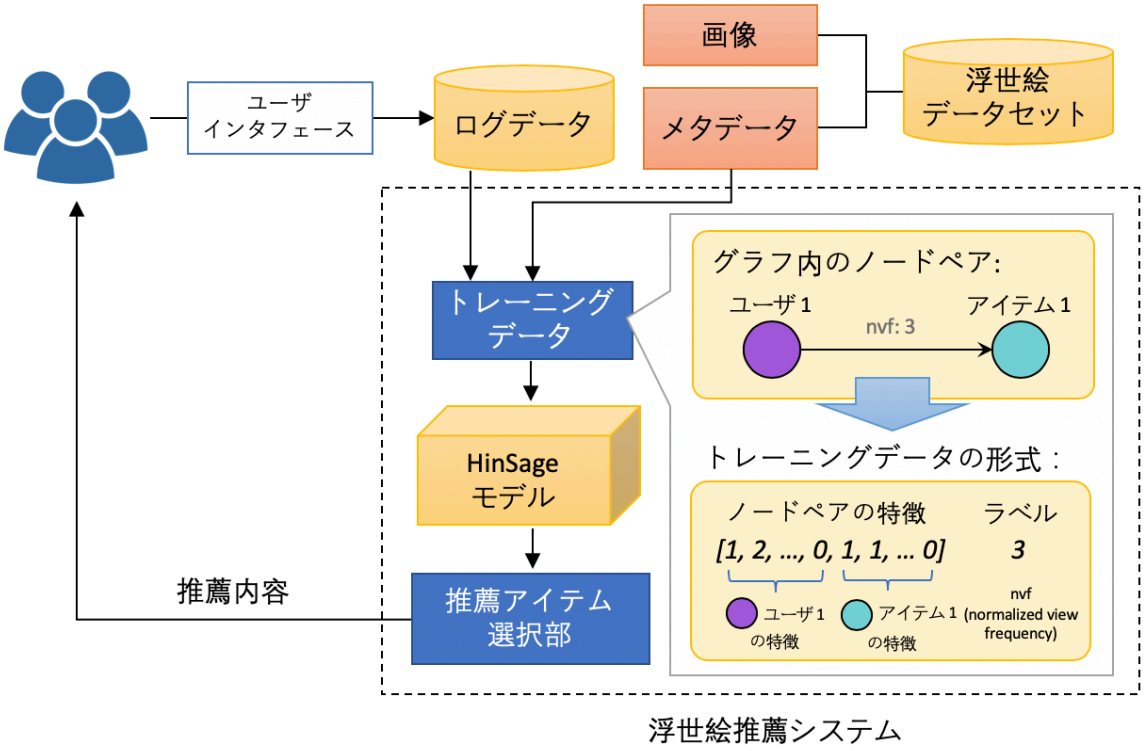


図 6　浮世絵推薦システムのフレームワーク
Figure 6　Ukiyo-e recommender system framework.

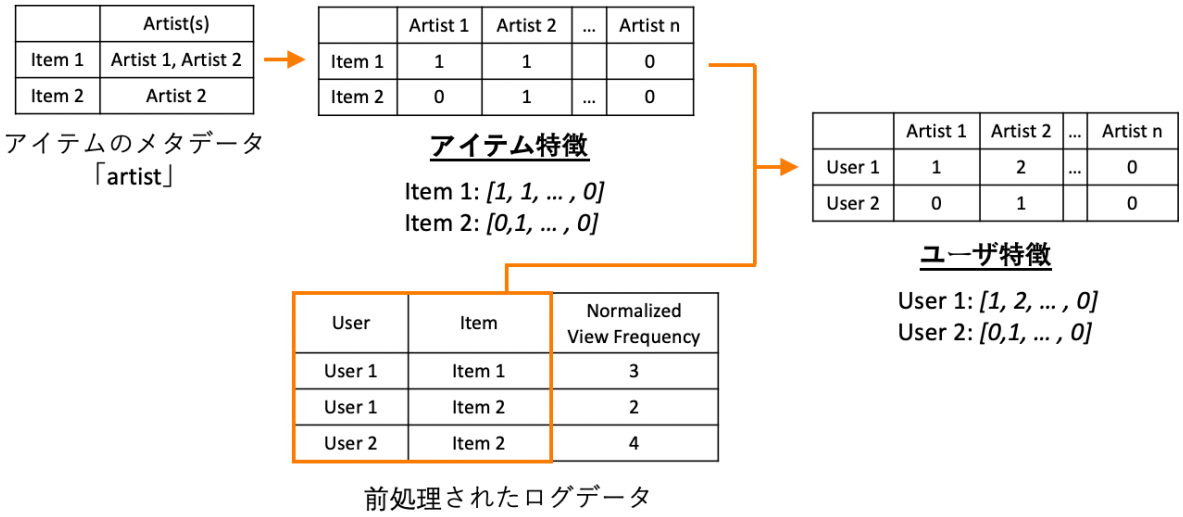


図 7　ユーザ特徴とアイテム特徴の生成
Figure 7　Generating user features and item features.

### ５.２アイテム特徴

本提案手法では，アイテムモデルは浮世絵のメタデータ「artist」を利用して構築される．図 7 の左上に示すように，元のメタデータはアイテムの絵師の名前（Artist 1，Artist 2）を記載している．そのデータを用いてアイテム特徴を one-hot エンコーディングで表す．例えば，Item 1 は Artist 1 と Artist 2 によって作成された場合，Item 1 のベクトルの対応する次元の値は「1」，他は「0」として示され，該当の浮世絵のベクトルは[1, 1, 0, …, 0]になる．

将来的に Wikidata の LOD データを利用して現在のデータセットを拡張することを考慮しているため，Wikidata に存在する絵師のみを利用する．合計で 130 人の絵師が存在するため，アイテムベクトルは全て 130 次元である．

### ５.３ユーザ特徴

4 章で述べた通り，編集者が編集モードなしの状態で Web ページをチェックする場合があり，その場合はあるユーザの閲覧頻度が極端に多くなる．また，ARC-UDB は正確に一人のユーザを判断するために必要なデータ（ログインデータなど）を収集していないため，ログデータの「IP」と「agent name」の両方でユーザを識別する方法を取る．したがって，できるだけ編集者や共用 PC などからのアクセスを削除するため，データを処理する際には，閲覧回数が 100 回以下のユーザのみを利用する．

さらに，ある程度アイテムを閲覧したユーザをトレーニングデータとして扱うという考えから，閲覧回数が 10 回以上のユーザのログデータを利用する．

ユーザモデルは，ログデータとメタデータの両方を利用して構築される．図 7 の中央の二つの表に表すように，ユーザ特徴はユーザが閲覧したアイテムの特徴の加算である．例えば，User 1 が Item 1 と Item 2 を閲覧した場合，User 1 のベクトルは Item 1 の特徴（[1, 1, 0, …, 0]）と Item 2 の特徴（[0, 1, 0, …, 0]）の加算で[1, 2, 0, …, 0]になる．ユーザ特徴もアイテム特徴と同様に全て 130 次元になる．

### ５.４HinSage モデル

本提案手法で利用するコアの深層学習モデルは，HinSage [5]と呼ばれる異質グラフのノードをベクトル化する表現学習モデルである．元の HinSage モデルは機械学習のタスクのためにベクトルの最適化を行うものである．

HinSage モデルに回帰層（link regression layer）を追加することにより，リンク予測のタスクに応用することができる[6]. 回帰層は，インプットしたベクトルの回帰タスクを行う．ここでノードペアの最適化した特徴がインプットであり，対応する予測 nvf が回帰層のアウトプットである．これにより，HinSage モデルを用いてリンク予測を行い，ユーザとアイテム間の関係を予測することができる．

## ６．予備実験

本提案手法の有効性を検証するため，予備実験を行った．実験で扱うデータは以下の通りである．

**ユーザデータ**：3,714 ユーザとその特徴
**アイテムデータ**：9,627 アイテム（浮世絵）とその特徴

**評価データ**: 33,623 ユーザ対アイテムの嗜好度レコード

モデルを訓練する段階では，ランダムに選んだ70%の評価データとそれに対応するユーザデータとアイテムデータがトレーニングデータであり，残りの 30%の評価データとそれに対応するユーザデータとアイテムデータがテストデータである．

提案手法を評価する指標として，以下の式で示す mean absolute error（MAE）と root mean square error（RMSE）を使用した．

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{\left(Y_i - \hat{Y}_i\right)^2}{n}}$$

各式中の$Y_i$が手法による出力の実際の値，$\hat{Y}_i$が予測値であり，二つの指標とも，$Y_i$と$\hat{Y}_i$の差を評価するものである．MAE と RMSE の値が低いほど，手法が優れていることを示す．

比較手法として，推薦システムの分野で一般的に利用される二つのアルゴリズム singular value decomposition（SVD）と item-to-item k-nearest neighbors（I2I-kNN）を使用する．

推薦システムにおいて，SVD は協調フィルタリングの一種であり，ユーザ対アイテムの評価履歴に基づいて，新たなユーザ対アイテムの評価を予測する方法である．以下の式のように，各アイテムは$q_i$で，各ユーザは$p_u$で表す．これらの 2 つのベクトルの内積が予測評価である．

$$expected\ rating = \hat{r}_{ui} = q_i^T p_u$$

以下の式のように，$q_i$と$p_u$は，その内積と既知の評価値との二乗誤差の差が最小になるように見つけることができる．

$$minimun(p,q)\sum_{u,i\in K}(r_{ui} - q_i^T \cdot p_u)^2$$

I2I-kNN は評価マトリックス内の各評価値を使用してアイテムのベクトルを表す．また，アイテムベクトル間のコサイン類似度を計算して推薦スコアを計算する．

予備実験の結果を表 1 に示す．

表 1　予備実験の結果
Table 1　Preliminary experimental results

| | 提案手法 | SVD | kNN |
|---|---|---|---|
| MAE | 0.4675 | 0.9089 | 0.9583 |
| RMSE | 0.6486 | 2.2499 | 2.3747 |

結果から，提案手法の誤差は I2I-kNN と SVD の誤差よりはるかに小さいことが分かった．特に，提案手法の二乗誤差を表す RMSE は比較的小さいことから，予測の外れ値が I2I-kNN や SVD より少ないことが示唆される．この結果は，我々が提案する手法が浮世絵データおよびグラフのデータ構造に適用可能であることを示していると考えられる．

ARC-UDB に保存されている浮世絵の数は非常に多く，事前のデータ分析により，ほとんどのユーザが一回だけしか ARC-UDB 内の浮世絵をアクセスしていないことがわかっている．そのため，ARC-UDB のデータは非常に疎であり，他の多くのデジタルアーカイブにも同様の問題があると考えられる．実験結果では，提案手法が既存の手法より優れていることを示しており，疎なデータに対する有効性を示唆している．さらに，今後 ARC-UDB のユーザおよび訪問数が増加するにつれて，データ量が増加する．これにより，トレーニングデータの量も増加し，モデルをより正確にすることができる．

## ６．あとがき

本論文では，ARC 浮世絵データに対してグラフベースの推薦システムを適用し，ユーザの嗜好を推測して推薦を行う手法を提案し，実験により有効性を示した．

現在の手法では，ユーザとアイテムを表すために絵師の名前のみを使用している．今後は，提案手法の有効性を引き続き最適化し，ユーザにより豊富な情報を推薦できるようにするため，ARC-UDB の他のメタデータや外部の LOD のデータなどを扱い，グラフ構造を充実させることを検討している．

## 参考文献

[1] http://www.dh-jac.net/db/nishikie/search_portal.php?enter=portal&lang=en
[2] Huang, Z., Chung, W., Ong, T. H., & Chen, H. “A graph-based recommender system for digital library.” In *Proceedings of the 2nd JCDH*. ACM, 2002.
[3] Aggarwal, C. C., Wolf, J. L., Wu, K. L., & Yu, P. S. “Horting hatches an egg: A new graph-theoretic approach to collaborative filtering.” In *Proceedings of the fifth ACM SIGKDD*, 1999.
[4] Musto, C., Lops, P., Basile, P., de Gemmis, M., & Semeraro, G. “Semantics-aware graph-based recommender systems exploiting linked open data.” In *Proceedings of the UMAP*. ACM, 2016.
[5] https://stellargraph.readthedocs.io/en/stable/api.html#module-stellargraph.layer.hinsage
[6] https://github.com/stellargraph/stellargraph/tree/develop/demos/link-prediction/hinsage

 