浮世絵ディジタルアーカイブのための作品関連性に基づく推薦システム

王 嘉韻[1]　Batjargal Biligsaikhan[2]　前田 亮[3]　川越 恭二[3]

1 立命館大学情報理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1
2 立命館大学衣笠総合研究機構 〒603-8577 京都市北区等持院北町 56-1
3 立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1
E-mail: 1 gr0278vx@ed.ritsumei.ac.jp

あらまし　浮世絵は日本の代表的な伝統芸術の一つであり、多くのディジタルアーカイブが浮世絵資料を保護・公開している。しかし、浮世絵ディジタルアーカイブの既存の検索方法はキーワード検索が中心であり、目的の画像を見つけるには専門知識が必要なため、一般ユーザにとっては使いにくい。本研究の目的は、浮世絵のメタデータをベクトル空間にマッピングし、ユーザが興味を持つ関連性のある浮世絵を見つけることである。本論文では、浮世絵の分類・著者・年代情報をベクトルで表し、作品間の類似度を計算し、利用者の嗜好に合った作品を見つけるための適切な推薦手法を検討する。

キーワード　推薦システム，ディジタルアーカイブ，テキスト処理

1. はじめに

浮世絵は日本の歴史を反映した貴重な芸術とみなされている。近年、浮世絵を保護・公開するために、浮世絵の絵画をデジタル化して保存するデジタルアーカイブが多く開発されている。「立命館大学アート・リサーチセンター」（ARC）の浮世絵ポータルデータベース[1]（以下、ARC-UP-DB という）は一つの例である。

デジタルアーカイブに保存された内容を容易に獲得するため、多くの浮世絵デジタルアーカイブには検索機能が装備されている。既存の浮世絵デジタルアーカイブにおける検索機能は、以下の二つのカテゴリに分類することができる：（ｉ）キーワード検索で画像を獲得する方法、（ii）一枚の画像をアップロードして、類似する画像を獲得する方法。この二つの情報検索の方法では、ユーザが自分の欲しいコンテンツを見つけるためには、ある程度浮世絵の専門知識を持っている必要がある。これらの機能は一部のユーザのニーズを満たすことができるが、キーワードや画像に詳しくない一般ユーザにとっては、既存の検索機能を利用して興味のある浮世絵を見つけることは困難である。

趣味や学習に浮世絵デジタルアーカイブを使用するようになった一般ユーザの増加を考慮して、我々は上記の問題を解決するための浮世絵推薦システム（Recommender System、以下 RS と言う）を提案する。RS は、ユーザが興味を持ちそうなコンテンツを予測するだけでなく、冗長な情報を排除することもできる。本研究は ARC-UP-DB のための推薦アルゴリズムを開発することに焦点を当てる。

ARC-UP-DB の推薦アルゴリズムの開発の難点としては、ユーザプロファイルと評価データは推薦結果を計算するための重要なデータであるが、ARC-UP-DB からはこの二つのデータが獲得できない。ARC-UP-DB にはログイン機能があるが、ただし、それの主な目的はユーザーの興味、行動などを収集することではない。そこで、我々は浮世絵自身が持つメタデータを利用し、内容ベースフィルタリング（Content-based filtering、以下 CBF と言う）に基づく方法で浮世絵間の類似度を計算し、推薦結果を獲得する。また、結果を検証するために、ユーザの実際のアクセスを表すログデータを用いて推薦結果の有効性を検証する。

我々のこれまでの研究[2]では、ARC-UP-DB のログデータを学習することでユーザの嗜好を予測する RBM 法を用いていた。これは数学的にユーザが興味のあるコンテンツの分布を推測する方法であり、ユーザが理解できる推薦理由についての説明が不足していた。そこで、今回は CBF を用いることで、推薦理由が推薦システムから得られるようになると考えられる。

2. 関連研究

デジタルアーカイブに対して推薦システムを実装した例の一つとして、Semeraro ら[3]の FIRSt（Folksonomy-based Item Recommender System）の提案が挙げられる。これは、ユーザからアイテムへの評価値と付与されたソーシャルタグを一般の CBF モデルに統合した。その結果、アイテムの説明の代わりにタグを使用するのではなく、タグによってユーザプロファイルを充実させると、推薦の精度が向上することが実験により示されている。

コンテンツを表すベクトルで類似するコンテンツを推薦する例の一つとして、Harry らが提案した

INTIMATE と呼ばれる、テキスト分類（text categorization）を使って映画のあらすじから推薦を行う推薦システムが挙げられる。INTIMATE は、bag of words の他にも、テキストをベクトル化する方法を試した。その結果、特徴ベースの手法（bag of words など）は、映画のあらすじのテキストのサイズがユーザ評価のサイズより大きい場合、通常の CBF モデルより優れていることを示している。

3. 提案手法の概要

3.1. 内容ベースフィルタリング（CBF）

CBF は、過去に行われたユーザの行動に基づいて推薦を行う[3]。また、推薦を目的としたオブジェクトのコンテンツ（テキスト、画像、音声など）を利用して推薦結果を生成する。そして、ユーザが購入したり、訪れた、聞いた、見た、あるいはランク付けしたアイテムに類似するアイテムを推薦する。metadata-based 法を用いた CBF は、コンテンツのメタデータを利用し、コンテンツ間の類似度を獲得する方法である。
本研究では浮世絵の三つのメタデータ、すなわち、分類・著者・年代情報を扱う。

本研究で扱うメタデータは 420 種の分類、269 名の著者と 152 種の年分のメタデータである。

3.2. Bag of Words

ベクトル化とは、テキスト文書の集合を数値の特徴ベクトルに変換する一般的なプロセスである。ベクトル化のため、トークン化、カウントおよび正規化を行う特定のテキスト処理手順は、Bag of Words（以下 BoW と言う）表現と呼ばれる。文書内の単語の相対位置情報は完全に無視され、文書は単語の出現によって記述される。この研究で扱うメタデータは、単語ごとに出現するので、単語の位置情報より、単語の出現頻度の方が重要だと考えられる。この考えで、BoW 法を利用し浮世絵をベクトル化する。ここでは、sklearn パッケージの CountVectorizer メソッドを使って BoW のベクトル化を行う。

3.3. ベクトル化

浮世絵をベクトル化するため、メタデータに出現した分類・著者・年代情報を列のタイトルに設定し、画像ごと該当するメタデータに「1」、該当しないメタデータに「0」を付与する。表 1 に簡単な例を示す。

表 1: メタデータをベクトル化する例

Asset_id	役者絵	風景画	見立絵	忠臣蔵
1	1	0	0	0
2	0	1	0	0
3	0	0	1	1

浮世絵の年代情報は二種類がある。一つは、一つの年分だけで表現する。例えば「弘化０４」と「1847」のような表記方法である。二つは、年代の期間で表現する。例えば「弘化０４～嘉永０５」のような表記方法である。

年代情報に対しては、まずは全ての情報を西暦に変換して、その後、二つのベクトル化方法を利用する。

一つ目（time 1）は、もし一つの年分だけで時間を表現場合、その年分を開始時間に記入し、もし年代期間がある場合、開始時間と終了時間両方も記入する。そして該当年分に時間のタグを付与する。表 2 には、六つの時間期間と時間タグの対応関係を示している。表 3 には「1847~1852」を表現する例を挙げる。省略した数字は全部「0」である。

表 2: 年代期間と時間タグの対応関係

期間	**1700~1749**	**1750~1799**	**1800~1849**
開始時間タグ	1	2	3
終了時間タグ	/	2	3
期間	**1850~1899**	**1900~1949**	**1950~**
開始時間タグ	4	5	6
終了時間タグ	4	5	/

表 3: time 1 ベクトル化方法の例

…	Start_3	…	End_4	…
…	1	…	1	…

二つ目（time 2）は、無論開始時間と終了時間、メタデータに出現した年分であれば「1」表示する。前文と同じ例を表 4 に挙げる。省略した数字は全て「0」である。

表 4: time 2 ベクトル化方法の例

…	1847	…	1852	…
…	1	…	1	…

「time 1」は一枚の浮世絵の時間を 10 次元のベクトルに、「time 2」は 152 次元にマッピングする。分類と著者のメタデータも含めると、前者は 699 次元のベクトル、後者は 841 次元のベクトルにマッピングする。

4. 評価実験

4.1. データセット

本研究で扱うデータは ARC-UP-DB の 2017 年 12 月 12 日から 2018 年 4 月 11 日までの ARC-DB に属する浮世絵（ARC-UP-DB は他の浮世絵データベースからのデータも格納するが、今回は研究対象としない）の 25,973 件のログデータと、そのログデータから獲得した全てのユーザがアクセスした ARC-UP-DB に属する 5648 枚の浮世絵のメタデータである。

最もユーザの興味を表すメタデータは浮世絵の分類・著者・年代情報と考えられるので、各々の浮世絵をその三つのメタデータによって表現する。有効な類似度を計算するために、少なくとも 1 つのメタデータが存在する 5,045 枚の浮世絵を選択して用いた。

4.2. 評価尺度

本研究で提案する二つの CBF 方法を検証するための評価尺度として、平均絶対誤差（MAE）(1)、精度（precision）(2)、および再現率（recall）(3) を使用する。e^tは予測と実データとの間の誤差である。

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|e^t| \quad (1)$$

$$precision = \frac{|\{Recommended\ contents\ that\ are\ relevant\}|}{|\{Recommended\ items\}|} \quad (2)$$

$$recall = \frac{|\{Recommended\ contents\ that\ are\ relevant\}|}{|\{Relevant\ items\}|} \quad (3)$$

4.3. 実験結果

本節では実験結果について述べる。表 5 に、10%から 90%まで、10%ずつ増加させた訓練データを使用した結果の平均値を示す。再現率と精度は、recall@all と precision@all で計算した。「time 1」は分類と著者のベクトルと time 1 のフォマットで表現する時間情報のベクトルと連結した方法をさす。「time 2」は時間のフォマットだけは「time 1」と違う。

表 5: 平均再現率、精度と MAE

ベクトル方法	再現率	精度	MAE
time 1	0.6180	0.0030	0.1437
time 2	0.6141	0.0050	0.2167

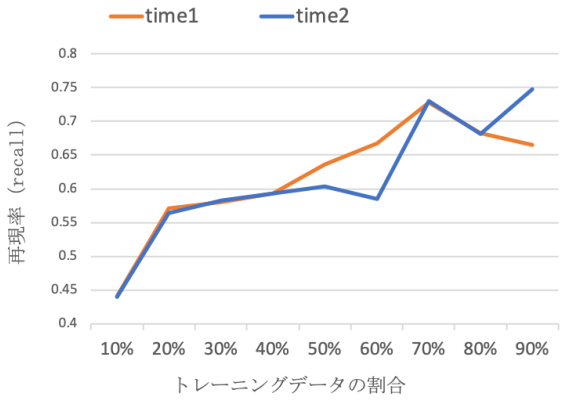

図 1: 再現率

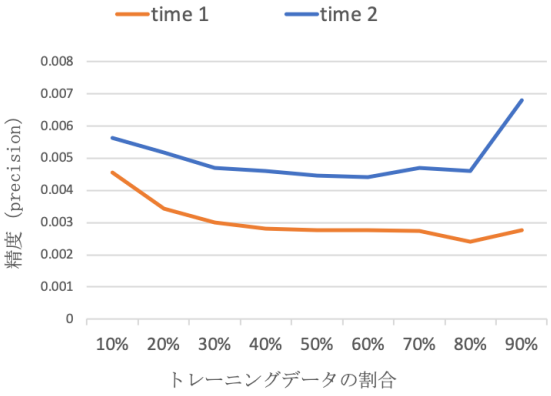

図 2: 精度

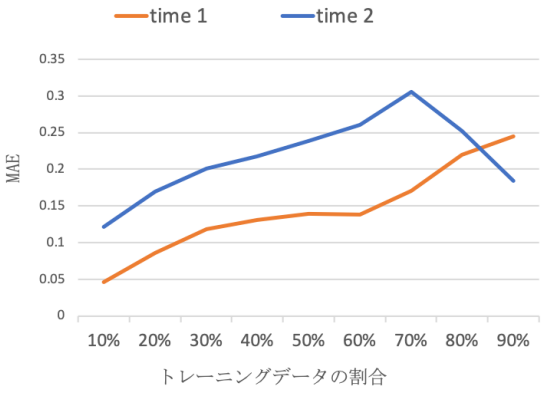

図 3: MAE

5. 実験結果の考察

今回は recall @ all と precision @ all を使用して結果を計算したが、これは精度に大きな影響を与えるが、再現率から、ある程度ユーザの嗜好を発見できていることが分かる。また、非常にスパースな行列であっても、この二つの方法の再現率は 0.5 以上である。これ

は、ユーザの嗜好とメタデータとは関連性があるためであると考えられる。さらに、time 1 のベクトル化法の精度は time 2 のベクトル化法よりも安定して高いことから、時間の分割方法は結果に影響を与えることが明らかにした。MAE 値のパフォーマンスはテストデータの割合が下がるほど増加しているが、ほぼ安定している。

6. おわりに

本研究では、ユーザが興味を持つ関連性のある浮世絵を見つけるため、浮世絵のメタデータを BoW と metadata-based の二つの方法でベクトル化して、類似するコンテンツを推薦する手法を提案した。今後は、浮世絵の類似度とユーザの興味の類似度の関連性を解明し、より精度を向上させることを目標とする。

参 考 文 献

[1] http://www.dh-jac.net/db/nishikie/search_portal.php

[2] Wang, Jiayun, and Kyoji Kawagoe. "Ukiyo-e recommender system using restricted boltzmann machine." Proceedings of the 19th International Conference on Information Integration and Web-based Applications & Services. ACM, 2017.

[3] Semeraro, Giovanni, et al. "A folksonomy-based recommender system for personalized access to digital artworks." Journal on Computing and Cultural Heritage (JOCCH) 5.3 (2012): 11.

[4] Mak, Harry, Irena Koprinska, and Josiah Poon. "Intimate: A web-based movie recommender using text categorization." Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on. IEEE, 2003.

[5] Bobadilla, Jesús, et al. "Recommender systems survey." Knowledge-based systems 46 (2013): 109-132.