ISSN 1346-2601 (Print)
ISSN 2436-5408 (Online) 2022 年 10 月 25 日公開

日本文化資源ディジタルアーカイブへの多言語情報アクセス技術

前田　亮（立命館大学情報理工学部　教授）
E-mail amaeda@is.ritsumei.ac.jp

1. 序論

本稿では、筆者が主宰する立命館大学情報理工学部ディジタル図書館研究室で現在進めている研究のうち、アート・リサーチセンター（以下 ARC）に関わる研究について、その概要を紹介する。

当研究室では、ARC が所蔵する資料を中心とする人文系の各種資料に対して、主にテキスト情報を活用して、情報検索や情報推薦、情報抽出などの情報アクセス技術を適用する研究を行っている。特に、海外の日本研究者から求められている多言語による情報アクセスの実現に力を入れて研究を進めてきた。最近では、テキスト情報に加え、画像情報など、異なる種類の情報を組み合わせて用いるマルチモーダル技術を適用し、より高度な情報アクセスを実現することを目指して研究を進めている。

本稿では、これらの研究のいくつかについて概要を紹介する。

2. 日本文化資源データベースのバイリンガル横断検索システム

本システムは、ARC が提供する各種日本文化資源のデータベースに対して、日英の二言語による横断検索を実現したものである [1)]。

本システムの特徴の一つは、事前にメタデータを収集するのではなく、検索時にリアルタイムで各データベースにアクセスし、検索結果を動的に統合して表示する点にある。これにより、データベースの更新に対してタイムラグが発生せず、常に最新の情報を検索することが可能となる。また、データベース毎に異なるメタデータ項目を Dublin Core に基づく主要なメタデータにマッピングすることで、検索結果の統合表示を容易にしている。

また、バイリンガル検索の機能として、専門用語辞書を用いた問合せ翻訳により、英語による問合せで日本語のメタデータを検索する言語横断検索の機能を実現している。さらに、日本語に不慣れな利用者による検索を容易にするために、日本語の形態素解析器および専門用語辞書により、日本語表記のメタデータからローマ字表記を自動生成して利用者に提示する機能を実現している。

本システムによる英語の検索結果の表示例を図 1 に示す。本システムにより、複数データベースを一つの問合せで一度に検索することができ、さらに日本語でメタデータが記述されたデータベースに対して日英二言語での検索が可能となる。本システムは、ARC ポータルデータベースの横断検索システム、ARC 所蔵資料公開データベースの横断検索システムとして、それぞれ公開している。
(https://www.arc.ritsumei.ac.jp/search/)

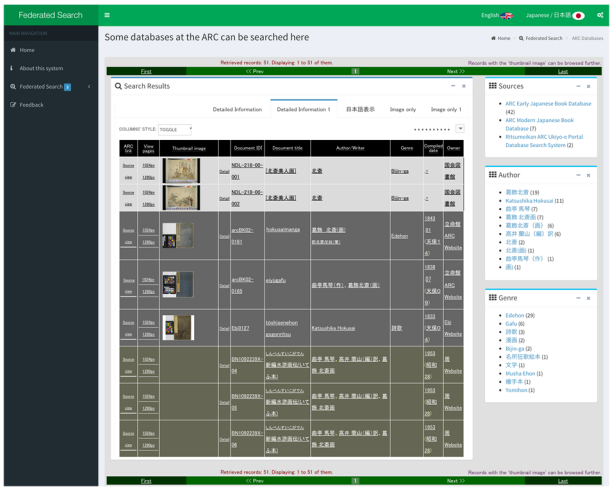
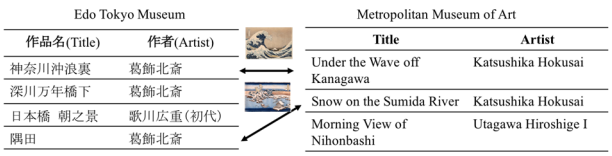


図 1 バイリンガル横断検索システムにおける英語の検索結果の表示例

3. 浮世絵データベースの言語横断レコード同定

本研究は、世界の美術館、博物館、研究機関等で公開されている浮世絵データベースを対象として、メタデータが記述されている言語が異なっていても、同一の浮世絵作品のレコードを発見することを目的としている。同一言語で記述されたメタデータを対象として同一レコードを発見する技術は、レコード同定技術として古くから研究されているが、異なる言語を対象とした言語横断レコード同定は、当研究室で考案した独自の技術である。浮世絵の言語横断レコード同定の例を図 2 に示す。



Edo Tokyo Museum			Metropolitan Museum of Art	
作品名(Title)	作者(Artist)		Title	Artist
神奈川沖浪裏	葛飾北斎	⟷	Under the Wave off Kanagawa	Katsushika Hokusai
深川万年橋下	葛飾北斎		Snow on the Sumida River	Katsushika Hokusai
日本橋 朝之景	歌川広重(初代)	↗	Morning View of Nihonbashi	Utagawa Hiroshige I
隅田	葛飾北斎			

図 2 浮世絵の言語横断レコード同定の例

言語横断レコード同定を実現するには、作品名などのメタデータ項目について、異なる言語間でのマッチングを行う必要があるが、図 2 の右側に見られるような海外の浮世絵データベースにおける翻訳された作品名と、日本語による元の作品名のマッチングは容易ではない。

本研究では、まず、固有名詞の抽出および翻字を行うことにより、マッチングにおいて重要となる固有名詞の抽出精度を向上する手法を提案している [2)]。

また、単語の意味のベクトル表現である単語分散表現を用いることで、メタデータ項目の意味的マッチングを行う手法を提案している [3)]。

さらに、各言語の単語分散表現におけるベクトル空間のマッピングを学習することにより、翻訳手法に一切依存せずに異なる言語間でのメタデータの類似度を測る手法を提案している [4)]。

同定精度にはまだ向上の余地があるものの、これらの手法により、これまで実現できなかった、異なる言語で記述された浮世絵レコードから同一作品を同定することが可能となった。

4. 役者評判記テキストからの役者情報の抽出

本研究では、江戸時代から明治初期まで発行された歌舞伎役者の芸評書である役者評判記について、役者評判記研究会 98 によってディジタルテキスト化されたものを対象として、テキスト中に含まれる役者に関する言及箇所を抽出する手法を提案している [5,6)]。役者評判記の原文の一部を図 3 に示す。

図 3 役者評判記『役者多名卸(江戸)』(出典:役者評判記『役者多名卸』3 項 立命館 ARC(BK04-0184))

具体的には、テキスト中から固有名詞等を抽出する手法である固有表現抽出の技術を用いて、機械学習により役者情報を抽出する。本研究 [6)]で抽出対象とする役者情報は、「役者名」「座本」「役種」「位付」「評価者」の 5 種類である。これらについて、モデルの学習に用いる正解データとして、単語ベースと文字ベースの 2 種類の正解データを作成した。単語ベースの正解データの作成には、形態素解析器 MeCab の辞書として近世口語(洒落本)UniDic を用いた。

深層学習に基づく固有表現抽出手法の一つである BiLSTM-CRF(Bidirectional Long Short-Term Memory -Conditional Random Fields)を用いた実験の結果、評価尺度の一つである F 値において、文字ベースの場合は 5 種類の役者情報の平均で約 87%、単語ベースの場合は平均で約 90%の精度が得られた。この結果より、古典資料のテキストから人物情報を自動的に抽出し、たとえば人物関係の分析や可視化に応用できる可能性を示すことができた。

5. 浮世絵中の落款印の認識および検索

浮世絵作品には、しばしば絵師の落款印が捺されていることがある。これらには文字だけでなく図案が含まれる場合もあり、その解読は必ずしも容易ではない。本研究では、肉筆の浮世絵を主な対象として、落款印全体の検索および落款印中の各文字の検索の実現を目指して研究を行っている [7,8)]。図 4 に、浮世絵の落款印全体を対象とした検索の例を示す。

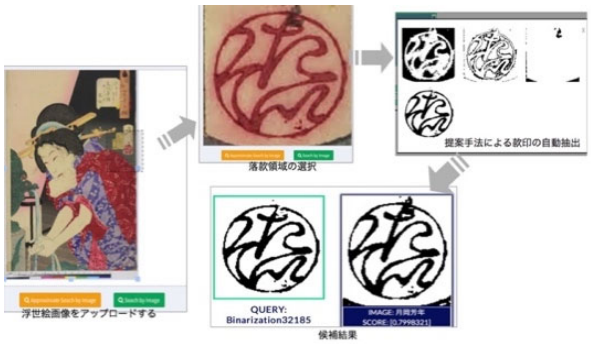


図 4 浮世絵の落款印全体を対象とした検索の例

落款印全体の検索では、落款印のデータベースとして「大日本書画名家大鑑〈落款印譜編〉, 荒木矩著, 第一書房(1934)」を用い、ARC 浮世絵ポータルデータベースに含まれる浮世絵の落款印を入力として検索を行う。

落款印中の各文字の検索では、まず文字領域の自動抽出を行い、次に抽出された各文字について文字認識を行う。落款印に含まれる篆文などの古代文字の認識は、各文字に対して単一あるいは少数の字形データしか入手できないことから、非常に困難なタスクであり、この実現は今後の課題である。

本研究の応用例として、得られた絵師の情報から Wikidata などのオープンデータを用いて浮世絵師の師弟関係を可視化することなどが考えられる。

6. 浮世絵データベースからの情報推薦

既存の人文系データベースでは、検索機能はほぼすべてに実装されているが、情報推薦の機能を実装しているものは少ない。本研究では、多様な浮世絵デー

タベースの利用者のニーズに応えるため、また、浮世絵データベースのさらなる有効活用を目指して、浮世絵データベースに対して情報推薦技術を適用する研究を行っている[9,10]。

本研究では、ARC 浮世絵ポータルデータベースのメタデータおよびアクセスログ(閲覧記録)データから、利用者と作品間の関係を示すグラフを構築し、これを基に推薦を行う。浮世絵データベースから構築したグラフの例を図 5 に示す。

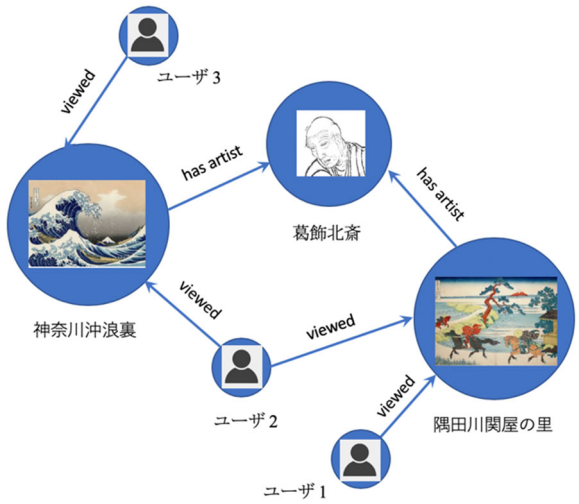


図 5 浮世絵データベースのアクセスログおよびメタデータから構築した、利用者と作品間の関係を表すグラフの例

構築したグラフから、リンク予測と呼ばれる手法を用いることで、利用者と作品間の潜在的な関係を推測することができる。これにより、ある利用者がある作品を好むかどうかを推測し、利用者が好むと思われる作品を推薦することが可能となる。本手法は、将来的には ARC 浮世絵ポータルデータベースの機能として搭載することを目標に研究を進めている。

7. まとめ

本稿では、現在当研究室で行っている、日本文化資源ディジタルアーカイブへの多言語情報アクセスの実現に向けたいくつかの研究について紹介した。

今後は、テキスト情報や画像情報など異なる種類の情報を組み合わせて用いるマルチモーダル技術を適用することで、より高度な情報アクセスを実現することを目指して研究を進め、ARC が提供する各種日本文化資源データベースのさらなる有効活用に繋げて行きたいと考えている。

［謝辞］

本稿で紹介した研究は、本学文学部の赤間亮教授からデータの提供およびアドバイスを得て進めてきたものです。また、4 章で紹介した役者評判記のディジタルテキストは、役者評判記研究会 98 から提供を受けたものです。ここに記して感謝の意を表します。

［注］

1) Biligsaikhan Batjargal, Akira Maeda, and Ryo Akama: Providing Bilingual Access to Multiple Japanese Humanities Databases: Text Retrieval Using English and Japanese Queries, In Jieh Hsiang, editor, *Digital Humanities: Between Past, Present, and Future*, pp. 351–367. National Taiwan University Press, 2016.

2) Yuting Song, Biligsaikhan Batjargal, and Akira Maeda: Recognition and Transliteration of Proper Nouns in Cross-Language Record Linkage by Constructing Transliterated Word Pairs, *International Journal of Asian Language Processing*, Vol. 27, No. 2, pp. 111–125, 2017.

3) Yuting Song, Biligsaikhan Batjargal, and Akira Maeda: Cross-Language Record Linkage based on Semantic Matching of Metadata, 日本データベース学会英文論文誌, Vol. 17, No. 1, pp. 1–8, 2019.

4) Yuting Song, Biligsaikhan Batjargal, and Akira Maeda: Title Matching for Finding Identical Metadata Records in Different Languages, In *Proceedings of the 13th International Conference on Metadata and Semantics Research (MTSR 2019)*, pp. 431–437, 2019.

5) 永井 規善, 前田 亮, 木村 文則, 赤間 亮:役者評判記からの人物表現抽出手法の提案, 人文科学とコンピュータシンポジウム論文集, pp. 145–150, 2014.

6) 川端 恵大, 前田 亮, 赤間 亮:役者評判記を用いた役者情報の抽出, 人文科学とコンピュータシンポジウム論文集, pp. 200–205, 2021.

7) 李 康穎, Biligsaikhan Batjargal, 前田 亮, 赤間 亮:落款印および関連情報の検索システムの構築:人物情報と人物関係ネットワークの自動抽出に向けて, 人文科学とコンピュータシンポジウム論文集, pp. 261–266, 2019.

8) Kangying Li, Biligsaikhan Batjargal, Akira Maeda, and Ryo Akama: Toward Exploring Artist Information from Seal Images in Ukiyo-e Collections, In *Conference Abstracts of Digital Humanities 2020*, 2020.

9) 王 嘉韻, Biligsaikhan Batjargal, 前田 亮, 川越 恭二, 赤間 亮: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, 人文科学とコンピュータシンポジウム論文集, pp. 165–170, 2019.

10) Jiayun Wang, Biligsaikhan Batjargal, Akira Maeda, Kyoji Kawagoe, and Ryo Akama: Making Ukiyo-e Easier to Discover: A Recommender System for Digital Archives, In *Conference Abstracts of Digital Humanities 2020*, 2020.