4ZL-04

画像とタイトルの類似性に基づく歴史資料探索支援システムの構築

田中 駿平 †　　奥野 拓 ‡
公立はこだて未来大学 †　公立はこだて未来大学 ‡

1 はじめに

デジタルアーカイブを利用する際，閲覧したい特定の歴史資料がなく，ブラウジングして興味のある資料を探索する場合がある．この場合，公開されている資料の全容を把握できていると，興味のある資料の発見が容易になる．しかし，膨大な資料が公開されていると，カテゴリやタグといった方法では大まかにしか絞り込めない場合がある．その場合，何回もページを遷移し，資料をブラウジングする必要があり，興味のある資料を発見することは容易ではない．そこで本研究では，デジタルアーカイブにおいて公開されている資料の全容の把握を容易にし，興味のある資料の発見を支援するシステムを構築する．類似資料を集約し，代表資料群として最初に提示することで，資料の全容を大まかに把握可能にする．そして，提示した代表資料から興味のあるものを選択すると，選択した資料と画像が類似する資料群，および選択した資料とタイトルが類似する資料群の一覧を提示する．

2 関連研究

画像資料を類似性によって分類し可視化する関連研究として，佐藤がニューラルネットワークを用いて類似性のある画像資料を近接するよう可視化するシステムを構築している [1]．このシステムでは，畳み込みニューラルネットワーク Inception-v3 [2] の識別層の手前にある第３プーリング層からの出力を，画像の特徴ベクトルとして抽出している．そして，抽出した特徴ベクトルを自己組織化マップで学習し，類似性のある画像資料が近くに配置されるように可視化を行っている．自己組織化マップとは，高次元データを任意の低次元空間（主に二次元）に写像するニューラルネットワークである．本研究では，同様の手法を用いて画像の特徴ベクトルの抽出および可視化を行う．

3 歴史資料探索支援システム

公開されている資料数が膨大な場合，全容の把握を容易にするためには，先に資料の全容を大まかに把握し，そこから段階的に資料を探索することが有効であると考えられる．そこで本研究では，最初に代表資料群を一覧で提示し，いずれかの資料を選択すると，選択した資料と画像が類似する資料群，および選択した資料とタイトルが類似する資料群へと段階的に探索するシステムを構築する．

本システムでは，2 と同様の手法により，資料の画像間の類似度の算出，および類似度の高い画像を持つ資料が近接するような配置を行う．これにより，似た特徴を持つ資料の把握を容易にする．しかし，画像の類似性によって配置するだけでは，画像間の類似性はないが関連がある資料が遠方に配置されることがある．そこで，選択された資料を基準に，資料に付与されたメタデータの類似性に基づく配置に並び替え可能にすることで，関連する資料の把握も容

Development of a Search Support System for Historical Records based on Images and Titles Similarity
† Shumpei Tanaka, Future University Hakodate
‡ Taku Okuno, Future University Hakodate

易にする．本研究では，メタデータのうち資料の内容を端的に表しているタイトルを用いる．詳細は 3.1 節で示す．

段階的に探索可能にするために，画像の類似性に基づく配置から，代表資料群を選出する．そして，最初に代表資料群を提示し，そこから類似資料へと探索可能にする．代表資料の選出手法の詳細は 3.2 で，提示手法は 3.3 で示す．

3.1 タイトル間の類似性に基づく配置

資料のタイトルに対して形態素解析を行い，名詞を抽出する．抽出した名詞から，Word2Vec の学習済みモデルである日本語 Wikipedia エンティティベクトル [3] を用いてベクトル表現を獲得し，タイトルの特徴ベクトルとする．1 つの資料タイトルから複数の名詞が抽出される場合は，抽出した名詞のベクトル表現の積を句ベクトルとし [4]，タイトルの特徴ベクトルとする．そして，各タイトル間のコサイン類似度を計算し，選択された資料を中心に類似度が高い資料から順に螺旋状に配置する．

3.2 代表資料の選出

資料の全容を大まかに把握可能にするために，代表資料群を選出する．選出手法として，まず，前述した画像の類似性に基づく配置結果を 3 × 3 の格子に分割し，格子内に含まれる画像の特徴ベクトルの平均ベクトルを算出する．そして，格子内において平均ベクトルに最も近い特徴ベクトルの画像を持つ資料を，格子内の代表資料として選出する．

3.3 ユーザへの提示

資料の全容を大まかに把握可能にするために，3.2 で選出した代表資料を提示する（図 1）．いずれかの資料が選択されると，資料の詳細説明と，画像の類似性による配置結果を基に選択した資料が中心となるよう配置した画面を提示する（図 2）．また，3.1 の結果を基に，資料の配置をタイトルの類似性による配置に並び替え可能にする．これらより，似た特徴を持つ資料の探索を支援する．

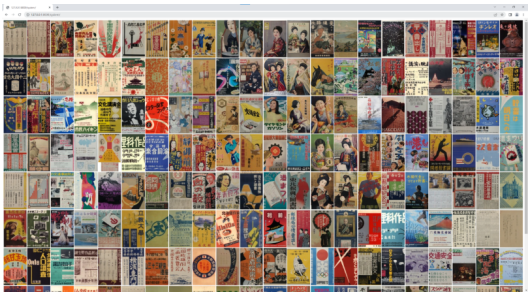

図 1　代表資料提示画面

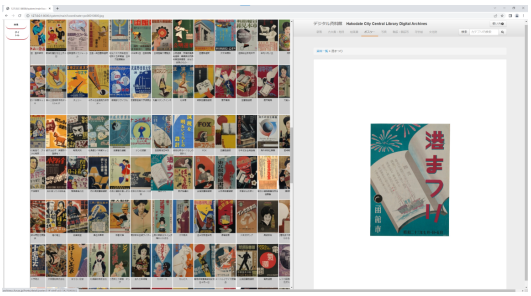

図 2　類似画像の資料の提示と資料の説明画面

4 まとめ

本稿では，デジタルアーカイブにおける歴史資料の探索を支援するシステムを提案した．今後は，既存システムとの比較実験を行い，本システムの有用性を評価する．

参考文献

[1] 佐藤太郎：DCNN Features による自己組織化マップを用いた歴史資料探索支援システムの構築，公立はこだて未来大学卒業論文（未公刊）(2019).

[2] Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the Inception Architecture for Computer Vision, *CVPR*, pp.2818-2826 (2016).

[3] 鈴木正敏, 松田耕史, 関根聡ほか：Wikipedia 記事に対する拡張固有表現ラベルの多重付与，言語処理学会第 22 回年次大会，pp.797-800 (2016).

[4] Mitchell, J. and Lapata, M.: Composition in Distributional Models of Semantics, Cognitive Science, Vol.34, pp.1388-1429 (2010).