

科学研究費助成事業 研究成果報告書

平成 28 年 6 月 12 日現在

機関番号：34315

研究種目：基盤研究(C) (一般)

研究期間：2012～2015

課題番号：24500300

研究課題名(和文)多言語デジタルアーカイブの統合検索に関する研究

研究課題名(英文)Research on Integrated Information Retrieval from Multilingual Digital Archives

研究代表者

前田 亮(Maeda, Akira)

立命館大学・情報理工学部・教授

研究者番号：20351322

交付決定額(研究期間全体)：(直接経費) 4,100,000円

研究成果の概要(和文)：本研究では、近年急速にデジタル化が進んでいる日本の古典史料の有効活用を目的として、世界中に散在する日本の古典史料のデジタルアーカイブに対して、統合的な情報アクセスを実現する技術を確立することを目標として研究を行った。その成果として、言語が異なるデジタルアーカイブの関連レコード同士を結びつけることで、これまで個別にアクセスするしか方法がなかった複数のデジタルアーカイブに対して統合的なアクセス手段を提供するという当初の目標を達成することができた。

研究成果の概要(英文)：In this research, we established integrated information access techniques for digital archives of Japanese cultural materials scattering around the world, with the objective of effectively utilizing such materials which are rapidly digitized in recent years. As a result, we achieved our research goal of providing integrated access means to multiple digital archives which could only be accessible individually, by automatically linking related records in such digital archives provided in different languages.

研究分野：図書館情報学

キーワード：メタデータ 言語横断レコード同定 言語横断エンティティリンクング 多言語処理 デジタル図書館 デジタルアーカイブ デジタルヒューマニティーズ

1. 研究開始当初の背景

近年、国内外の図書館・博物館・美術館・文書館などにおいて、資料のデジタル化および公開が進んでいる。これらは、通常個々の機関が個別にデジタルアーカイブとして公開を行っており、アクセスも個別に行う必要がある。

国立国会図書館デジタルアーカイブポータル PORTA や、人間文化研究機構における研究資源共有化システムなど、統合アクセスを実現しているものもあるが、これらは基本的に、事前に統合検索システム側でメタデータを収集する「ハーベスティング」もしくは横断検索のための登録作業が必要であり、いずれにしても人的コストがかかるのが現状である。

また、これらは基本的に日本語で提供されているデジタルアーカイブを対象としているが、たとえば浮世絵などは、大英博物館、ヴィクトリア・アンド・アルバート博物館、ボストン美術館、米国議会図書館など、海外のさまざまな機関でデジタル化され、公開されている。現状では、これらのデジタルアーカイブを統合的に検索するシステムは存在しない。

複数デジタルアーカイブの統合検索に関しては、日本では人間文化研究機構を中心とする研究資源共有化プロジェクトにおいて先駆的な研究が行われ、研究資源共有化システムとして公開されている。また、国立国会図書館のデジタルアーカイブポータルは、国内のさまざまな機関が提供するデジタルアーカイブの統合検索を実現している。海外では、Europeana や HathiTrust などが大規模なデジタルアーカイブ統合検索システムを提供している。

2. 研究の目的

本研究では、近年急速にデジタル化が進んでいる日本の古典史料の有効活用を目的として、世界中に散在する日本の古典史料のデジタルアーカイブに対して、統合的な情報アクセスを実現する技術を開発することを目指した。

これにより、これまで個別にアクセスするしか方法がなかった複数のデジタルアーカイブに対して統合的なアクセスが可能となり、国内外における日本文化研究に大いに貢献することが期待される。

本研究では、デジタルアーカイブの統合検索システムの検索対象となりにくい小規模なデジタルアーカイブや、英語など日本語以外の言語で提供されている日本の古典史料を含む、多言語からなる複数のデジタルアーカイブの統合検索を可能とするシステムを構築することを目指した。

3. 研究の方法

本研究では、多言語デジタルアーカイブの統合検索の実現に向けて、主に以下の研究を行った。

- 複数デジタルアーカイブにおける作者の人物同定手法の開発
- 複数デジタルアーカイブにおける言語横断レコード同定手法の開発
- 複数デジタルアーカイブに対する言語横断エンティティリンク手法の開発

(1) 複数デジタルアーカイブにおける作者の人物同定手法の開発

本研究では、複数デジタルアーカイブにおける同一作者のレコード間のリンクの実現の検討を行った。

具体的には、浮世絵の作者（絵師）の情報を対象とし、国立国会図書館が提供している典拠データ検索・提供サービス（Web NDL Authorities）および、主に欧米の国立図書館の典拠データ間をリンクした仮想的な典拠データである VIAF（Virtual International Authority File）における著者名典拠に結び付ける手法を検討した。

このうち、特に国立国会図書館の典拠データには、浮世絵師の情報が多く含まれており、また Web NDL Authorities と VIAF のいずれも SPARQL（SPARQL Protocol and RDF Query Language）による検索に対応しているため、Linked Data 化が容易である。

また、Web NDL Authorities では、人名の読みの情報からローマ字表記に変換することも可能である。ただし、浮世絵師などの場合、同一人物であっても襲名により名前が変わる場合や、逆に同名の人物が複数存在する場合があります。これらの人物の同定が必要である。そのため、他のメタデータ項目に含まれる年代の情報などを利用して人物の同定を行う手法を検討した。

(2) 複数デジタルアーカイブにおける言語横断レコード同定手法の開発

本研究では、言語が異なる複数デジタルアーカイブから同一作品を自動的に同定する手法について研究を行った。

浮世絵は木版画であるため、同一作品が複数のデータベースに所蔵されていることが多くあるが、メタデータスキーマや記述言語の違いから、同一作品を見つけ出すことは容易ではない。そこで本研究では、メタデータの特定の項目（作品の題名など）を用い、題名の音訳や英訳など、表記や言語が異なる場合であっても同一作品を自動的に見つけ出すための手法を開発した。

具体的には、比較対象のデータベースの全レコードから、まず浮世絵の作者（絵師）のメタデータを用いて同一作者に絞り込みを

行う。次に、検索対象の作品名と絞り込み後の対象レコードの作品名の類似度を求め、一定以上の類似度を持つレコードを同一作品と推定する。作品名の類似度の計算は、作品名の表記や言語の違いによって以下のような手法を使い分ける。

同言語・同表記の作品名同士の比較：文字列の比較に良く用いられる編集距離(レーベンシュタイン距離)を用いて類似度を求める。

作品名の音訳(ローマ字表記)と英訳の比較：作品名中の固有名詞に着目し、固有名詞の一致数および対訳辞書による訳語一致数を元に類似度を求める。

作品の原題と音訳の比較：原題に対し形態素解析器などを用いて読みを抽出し、それをローマ字化したものと音訳を比較する。その際、単語ごとの編集距離および語順の情報を用いて類似度を計算する。

作品の原題と英訳の比較：まず、原題を辞書との最長一致により単語に分割し、逐語訳を行う。その逐語訳と作品名の英訳に対して、固有名詞および一般名詞の一致数を基に類似度を算出する。

(3) 複数デジタルアーカイブに対する言語横断エンティティリンクング手法の開発

本研究では、デジタルアーカイブ内のメタデータなどのテキスト中で言及されているエンティティ(実体)から、それを説明する別言語のデジタルアーカイブのレコードに自動的にリンクする言語横断エンティティリンクングの研究を行った。

具体的には、リンク対象とする人名などの固有表現の自動抽出手法、エンティティ記述の翻訳手法、対象リンク先の曖昧性解消手法について研究を行った。

固有表現の自動抽出については、SVM(Support Vector Machine)による機械学習を用い、文字単位で学習することで、古典資料など形態素解析が困難な日本語テキストから固有表現を抽出する手法を確立した。

エンティティ記述の翻訳については、複数の翻訳手法による翻訳結果と文字列の部分一致を組み合わせて用いることにより、訳語抽出の再現率向上を図った。

対象リンク先の曖昧性解消手法については、リンク先の候補となる各レコードに含まれるテキストと、リンク元となるテキストの内容の類似度を計算することにより、適切なリンク先に絞り込む手法を提案した。

4. 研究成果

(1) 複数デジタルアーカイブにおける作者の人物同定手法の開発

本研究で提案した、複数デジタルアーカイブにおける同一作者のレコード間のリンクのための人物同定の精度の定量的評価を行った。その結果、姓と名の両方が含まれる人名の場合で約99.8%と、非常に高い精度で人物を同定することができた。

これにより、本研究の目標である多言語デジタルアーカイブの統合検索の実現に向けて一定の見通しが得られた。

(2) 複数デジタルアーカイブにおける言語横断レコード同定手法の開発

本研究で提案した、浮世絵の作品名の音訳と英訳を用いた同一作品の同定精度の実験を行った結果、MAP(Mean Average Precision)において81.4%の精度が得られた。

また、浮世絵の作品名の原題と英訳に対して提案手法による同一作品の同定精度の実験を行った結果、約78%の正解率が得られた。

この提案手法と、(1)で提案した同一作者レコードの同定手法を組み合わせることにより、異言語の複数デジタルアーカイブから、高い精度で同一作者・同一作品レコードを抽出できることを示した。

(3) 複数デジタルアーカイブに対する言語横断エンティティリンクング手法の開発

本研究で提案した固有表現の自動抽出手法に関して、平安時代後期に書かれた日本語古典資料からの固有表現抽出の評価実験を行った結果、再現率で約72%、適合率で約92%が得られた。

また、エンティティ記述の翻訳手法に関して、日本語の新聞記事から中国語のWikipedia記事にリンクする実験を行った結果、約90%の再現率が得られた。

対象リンク先の曖昧性解消手法に関しては、日本語の新聞記事から中国語のオンライン百科事典「百度百科」の記事にリンクする実験を行った結果、約97%の正解率が得られた。

これらの実験結果より、単に複数デジタルアーカイブの同一レコード間をリンクするだけでなく、デジタルアーカイブのレコード中の作品の説明文などのテキストから、関連するデジタルアーカイブのレコードへのリンクが実現できることを示した。

上記の三つの研究成果により、言語が異なるデジタルアーカイブの関連レコード同士を結びつけることで、これまで個別にアクセスするしか方法がなかった複数のデジタルアーカイブに対して統一的なアクセス手段を提供するという当初の目標を達成することができた。

5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者に

は下線)

〔雑誌論文〕(計6件)

1. Kensuke Horita, Fuminori Kimura, and Akira Maeda. Automatic Keyword Extraction for Wikification of East Asian Language Documents. *International Journal of Computer Theory and Engineering*, Vol. 8, No. 1, pp. 32-35, Feb. 2016. 査読有 DOI:10.7763/IJCTE.2016.V8.1015
2. Takafumi Sato, Makoto Goto, Fuminori Kimura, and Akira Maeda. Developing a Collaborative Annotation System for Historical Documents by Multiple Humanities Researchers. *International Journal of Computer Theory and Engineering*, Vol. 8, No. 1, pp. 88-93, Feb. 2016. 査読有 DOI:10.7763/IJCTE.2016.V8.1025
3. Fuminori Kimura, Takahiko Osaki, Taro Tezuka, and Akira Maeda. Visualization of relationships among historical persons from Japanese historical documents. *Literary and Linguistic Computing*, Vol. 28, No. 2, pp. 271-278, Jun. 2013. 査読有 DOI: 10.1093/lc/fqs045
4. Biligsaikhan Batjargal, Takeo Kuyama, Fuminori Kimura, and Akira Maeda. Linked data driven multilingual access to diverse Japanese Ukiyo-e databases by generating links dynamically. *Literary and Linguistic Computing*, Vol. 28, No. 4, pp. 522-530, Dec. 2013. 査読有 DOI: 10.1093/lc/ftq058
5. Fuminori Kimura, Hiroshi Urae, Taro Tezuka, and Akira Maeda. Multilingual Translation Support for Web Pages Using Structural and Semantic Analysis. *IAENG International Journal of Computer Science*, Vol. 39, No. 3, pp. 276-285, Aug. 2012. 査読有 http://www.iaeng.org/IJCS/issues_v39/issue_3/IJCS_39_3_07.pdf
6. Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, Fuminori Kimura, and Akira Maeda. Developing a Digital Library of Historical Records in Traditional Mongolian Script. *International Journal of Digital Library Systems*, Vol. 3, No. 1, pp. 33-52, Jun. 2012. 査読有 DOI:10.4018/jdls.2012010103

〔学会発表〕(計37件)

1. Xiang Song, Jialiang Zhou, Fuminori Kimura, and Akira Maeda. A Japanese-Chinese Cross-Language Entity Linking Method with Entity Disambiguation Based on Document Similarity. In *Proceedings of the 2nd International Conference on Knowledge*

(ICK 2016), 2016年3月18日, パリ(フランス)

2. Yuting Song, Taisuke Kimura, Biligsaikhan Batjargal, and Akira Maeda. An Approach to Build a Proper Noun Dictionary for Record Linkage across Humanities Databases in Different Languages. 第8回データ工学と情報マネジメントに関するフォーラム (DEIM2016) 論文集, 2016年2月29日, ヒルトン福岡シーホーク(福岡市)
3. 浦田 智昭, 前田 亮. 音楽記事中のアーティスト名を対象としたエンティティリンクング. 第8回データ工学と情報マネジメントに関するフォーラム (DEIM2016) 論文集, 2016年2月29日, ヒルトン福岡シーホーク(福岡市)
4. 木村 泰典, Biligsaikhan Batjargal, 木村 文則, 前田 亮. 多言語の浮世絵データベース間における同一作品の同定手法の提案. *人文科学とコンピュータシンポジウム論文集*, pp. 117-124, 2015年12月19日, 同志社大学京田辺キャンパス(京田辺市)
5. 佐藤 貴文, 後藤 真, 木村 文則, 前田 亮. 歴史資料からの人文系研究者への注釈候補の提示手法の構築. *人文科学とコンピュータシンポジウム論文集*, pp. 165-172, 2015年12月19日, 同志社大学京田辺キャンパス(京田辺市)
6. Biligsaikhan Batjargal, Akira Maeda, and Ryo Akama. Providing Bilingual Access to Multiple Japanese Humanities Databases: Text Retrieval Using English and Japanese Queries. In *Proceedings of the 6th International Conference of Digital Archives and Digital Humanities (DADH2015)*, pp. 431-442, 2015年12月1日, 台北(台湾)
7. Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, and Akira Maeda. Personal Name Extraction from Mongolian Historical Documents Using Machine Learning. In *Proceedings of the 6th International Conference of Digital Archives and Digital Humanities (DADH2015)*, pp. 419-430, 2015年12月1日, 台北(台湾)
8. Noriyoshi Nagai, Fuminori Kimura, Akira Maeda, and Ryo Akama. Personal Name Extraction from Japanese Historical Documents Using Machine Learning. In *Proceedings of the International Conference on Culture and Computing (Culture and Computing 2015)*, pp. 207-208, 2015年10月18日, 京都大学(京都市)
9. Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, and Akira Maeda. An Approach to Named Entity Extraction from Mongolian Historical Documents. In

- Proceedings of the International Conference on Culture and Computing (Culture and Computing 2015), pp. 205-206, 2015年10月18日, 京都大学(京都市)
10. Xiang Song, Jialiang Zhou, Fuminori Kimura, and Akira Maeda. A Japanese-Chinese Cross-Language Entity Linking Method Based on Appropriateness of Term Description Pages. In Proceedings of the 4th IIAI International Congress on Advanced Applied Informatics (IIAI AAI 2015), pp. 234-238, 2015年7月12日, 岡山コンベンションセンター(岡山市)
 11. Taisuke Kimura, Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda. Finding the Same Artworks from Multiple Databases in Different Languages. In Conference Abstracts of Digital Humanities 2015, 2015年7月1日, シドニー(オーストラリア)
 12. Takafumi Sato, Makoto Goto, Fuminori Kimura, and Akira Maeda. Extracting Key Phrases for Suggesting Annotation Candidates from Japanese Historical Document. In Conference Abstracts of Digital Humanities 2015, 2015年7月1日, シドニー(オーストラリア)
 13. Jialiang Zhou, Xiang Song, Fuminori Kimura, and Akira Maeda. A Cross-Language Entity Linking Method Using Combination of Multiple Translation Methods. In Proceedings of the 4th ICT International Student Project Conference (ICT-ISPC2015), 2015年5月23日, 東京農工大学(小金井市)
 14. 周 佳良, 宋 翔, 堀田 健介, 木村 文則, 前田 亮. オンライン百科事典を対象とした日中言語間エンティティリンクング手法の提案 - 日本語文章中の重要語の翻訳手法 -. 第77回情報処理学会全国大会講演論文集, 第1分冊(3N-01) pp. 629-630, 2015年3月18日, 京都大学吉田キャンパス(京都市)
 15. 宋 翔, 周 佳良, 堀田 健介, 木村 文則, 前田 亮. オンライン百科事典を対象とした日中言語間エンティティリンクング手法の提案 - 適切な用語説明ページの抽出手法 -. 第77回情報処理学会全国大会講演論文集, 第1分冊(3N-02) pp. 631-632, 2015年3月18日, 京都大学吉田キャンパス(京都市)
 16. 木村 泰典, Biligsaikhan Batjargal, 木村 文則, 前田 亮. 言語が異なる浮世絵データベース間における同一作品の同定手法の提案. 第77回情報処理学会全国大会講演論文集, 第4分冊(1ZC-05), pp. 639-640, 2015年3月17日, 京都大学吉田キャンパス(京都市)
 17. 原田 一慧, 木村 文則, 前田 亮. Wikipedia 記事の言語間差異抽出手法の提案. 第7回データ工学と情報マネジメントに関するフォーラム(DEIM2015)論文集, 2015年3月4日, 磐梯熱海ホテル華の湯(郡山市)
 18. 永井 規善, 木村 文則, 前田 亮, 赤間 亮. 役者評判記からの人物に関する表現の自動抽出手法. 第4回知識・芸術・文化情報学研究会, 2015年2月7日, 立命館大学大阪梅田キャンパス(大阪市)
 19. 佐藤 貴文, 後藤 真, 木村 文則, 前田 亮. 東大寺要録からの歴史知識情報の抽出 - 注釈情報の共有を目指して -. 人文科学とコンピュータシンポジウム論文集, pp. 93-100, 2014年12月13日, 国立情報学研究所(東京都)
 20. 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年12月13日, 国立情報学研究所(東京都)
 21. Biligsaikhan Batjargal, Takeo Kuyama, Fuminori Kimura, and Akira Maeda. Identifying the Same Records across multiple Ukiyo-e Image Databases Using Textual Data in Different Languages. In Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014), pp. 193-196, 2014年9月10日, ロンドン(イギリス)
 22. Fuminori Kimura and Akira Maeda. Method for Supporting Analysis of Personal Relationships through Place Names Extracted from Documents. In Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014), pp. 253-256, 2014年9月10日, ロンドン(イギリス)
 23. Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, Fuminori Kimura, and Akira Maeda. An Approach to Named Entity Extraction from Historical Documents in Traditional Mongolian Script. In Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014), 2014年9月9日, ロンドン(イギリス)
 24. 加藤 拓磨, 久山 岳夫, Biligsaikhan Batjargal, 木村 文則, 前田 亮. メタデータを用いた異言語浮世絵データベース間における同一作品の同定手法. 第3回知識・芸術・文化情報学研究会, 2014年2月8日, 立命館大学大阪梅田キャンパス(大阪市)
 25. 佐藤 貴文, 後藤 真, 木村 文則, 前田 亮. 複数の人文系研究者による史料注釈を可能とするWebシステムの試作—『東大寺要録』を用いて—. 人文科学とコンピュータシンポジウム論文集, pp. 57-64, 2013年12月12日, 京都大学吉田キャンパス(京都市)
 26. 久山 岳夫, Biligsaikhan Batjargal, 木村

- 文則, 前田 亮. 複数の異種浮世絵データベース間における同一作品の同定手法の提案. 人文科学とコンピュータシンポジウム論文集, pp. 225-232, 2013年12月14日, 京都大学吉田キャンパス(京都市)
27. 吉村 衛, 木村 文則, 前田 亮. 古文テキストからの人物表現抽出. 人文科学とコンピュータシンポジウム論文集, pp. 97-102, 2013年12月13日, 京都大学吉田キャンパス(京都市)
28. Akira Maeda. Multilingual Access to Diverse Digital Libraries and Archives: A Linked Data Approach. Invited talk at the Fourth International Conference on Digital Libraries (ICDL2013), 2013年11月28日, ニューデリー(インド)
29. Biligsaikhan Batjargal, 木村 文則, 前田 亮. 浮世絵を対象とした多言語・異種データベースの横断検索. 第19回公開シンポジウム「人文科学とデータベース」論文集, pp. 27-32, 2013年11月30日, 立命館大学衣笠キャンパス(京都市)
30. Fuminori Kimura, Katsuhiko Mitsui, and Akira Maeda. Extraction of Linked Data Triples from Japanese Wikipedia Text of Ukiyo-e Painters. In Proceedings of the International Conference on Culture and Computing (Culture and Computing 2013), pp. 192-193, 2013年9月17日, 立命館大学朱雀キャンパス(京都市)
31. Biligsaikhan Batjargal, Fuminori Kimura, Garmaabazar Khaltarkhuu, and Akira Maeda. Applying Text Encoding Initiative Guidelines to a Historical Record in Traditional Mongolian Script. In Proceedings of the International Conference on Culture and Computing (Culture and Computing 2013), pp. 141-142, Kyoto, Japan, 2013年9月16日, 立命館大学朱雀キャンパス(京都市)
32. Biligsaikhan Batjargal, Takeo Kuyama, Fuminori Kimura, and Akira Maeda. Linked Data Driven Dynamic Web Services for Providing Multilingual Access to Diverse Japanese Humanities Databases. In Proceedings of the 13th International Conference on Dublin Core and Metadata Applications (DC-2013), pp. 19-24, 2013年9月3日, リスボン(ポルトガル)
33. Mamoru Yoshimura, Fuminori Kimura, and Akira Maeda. Personal Name Extraction from Ancient Japanese Texts. In Proceedings of the Exploration, Navigation and Retrieval of Information in Cultural Heritage ENRICH 2013 Workshop, pp. 31-34, Dublin, Ireland, 2013年8月1日, ダブリン(アイルランド)
34. Biligsaikhan Batjargal, Takeo Kuyama, Fuminori Kimura, and Akira Maeda. A Linked Data Driven Approach on Cross Language Information Access to Diverse Japanese Databases. In Book of Abstracts of the 5th International Conference on Qualitative and Quantitative Methods in Libraries (QQML2013), pp. 178-179, 2013年6月7日, ローマ(イタリア)
35. 久山 岳夫, Biligsaikhan Batjargal, 木村 文則, 前田 亮. 動的リンク生成による浮世絵データベース間の多言語統合アクセス手法の提案. 人文科学とコンピュータシンポジウム論文集, pp. 231-238, 2012年11月17日, 北海道大学札幌キャンパス(札幌市)
36. Mamoru Yoshimura, Fuminori Kimura, and Akira Maeda. Word Segmentation for Text in Japanese Ancient Writings Based on Probability of Character N-grams. In Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries (ICADL2012), pp. 313-316, 2012年11月14日, 台北(台湾)
37. Takeo Kuyama, Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda. Integrated Multilingual Access to Diverse Japanese Humanities Digital Archives by Dynamically Linking Data. In Conference Abstracts of Digital Humanities 2012, pp. 473-476, 2012年7月18日, ハンブルク(ドイツ)

6. 研究組織

(1) 研究代表者

前田 亮 (MAEDA, Akira)
立命館大学・情報理工学部・教授
研究者番号: 20351322

(2) 研究分担者

木村 文則 (KIMURA, Fuminori)
尾道市立大学・経済情報学部・講師
研究者番号: 70516690