

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 5 月 31 日現在

機関番号：34315

研究種目：若手研究（B）

研究期間：2009 ～ 2011

課題番号：21700271

研究課題名（和文） 言語・時代・文化横断型の情報アクセスに関する研究

研究課題名（英文） Studies on Information Access across Languages, Periods, and Cultures

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

研究成果の概要（和文）：

本研究では、近年急速に電子化が進んでいる過去の古い資料の有効活用を目的に、言語だけではなく時代や文化をも横断する情報アクセス技術確立することを目標として研究を行った。具体的には「日本語古典史料のテキスト処理手法の開発」、「人文系データベースの多言語統合検索システム」、「伝統的モンゴル文字文書のデジタル図書館システム」、「多言語 Web ページの作成支援システム」、「確率モデルに基づく情報推薦手法」に関する研究を行った。

研究成果の概要（英文）：

In this research, we established information access techniques across languages, periods, and cultures, with the aim of promoting effective use of digitized historical documents, which are rapidly increasing in recent years. Specifically, we developed “text processing techniques for ancient Japanese historical materials”, “multilingual integrated search system for humanities databases”, “digital library of documents in traditional Mongolian script”, “system for supporting creation of multilingual web pages”, and “information recommendation techniques based on probabilistic models”.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合計 |
|---------|-----------|---------|-----------|
| 2009 年度 | 800,000 | 240,000 | 1,040,000 |
| 2010 年度 | 800,000 | 240,000 | 1,040,000 |
| 2011 年度 | 800,000 | 240,000 | 1,040,000 |
| 年度 | | | |
| 年度 | | | |
| 総計 | 2,400,000 | 720,000 | 3,120,000 |

研究分野：図書館情報学

科研費の分科・細目：情報学，図書館情報学・人文社会情報学

キーワード：情報検索，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，テキストマイニング，可視化

1. 研究開始当初の背景

現在の Web 上では膨大な量の情報が提供されているが、その多くは、Web が登場した後、つまりここ数十年程度の間で作られたものである。これに対して、紙媒体を対象とす

る図書館は、紀元前から現在まで人類が残した膨大な知識を蓄積しており、近年に作られた情報がほとんどである現在の Web とは、蓄積されている情報のタイムスパンにおいて大きな差がある。また、情報の量においても、Web 上の索引付け可能な情報は 2005 年の時

点で約 115 億ページ、約 150 テラバイトであるが、紙媒体の図書館にはそれ以上の量（米
国議会図書館の蔵書だけで約 2,000 万冊、約
160 テラバイト）の情報が蓄積されていると
言われている（M. Lesk, “Understanding Digital
Libraries”, 2nd ed., 2005）。

近年、インターネット関連企業・政府機
関・大学・図書館などの協力による図書館蔵
書の大規模な電子化プロジェクト（たとえば
Google Book Search, Open Content Alliance,
World Digital Library, HathiTrust など）が多数
立ち上がっており、従来は紙媒体しかないと
特定図書館でしか閲覧できなかった古代から
現代にかけての膨大な量の知識の蓄積に対
して Web 上で容易にアクセス可能になり
つつある。

2. 研究の目的

本研究では、近年急速に電子化が進んで
いる過去の古い資料の有効活用を目的に、従
来から研究されている言語横断情報検索
（Cross-Language Information Retrieval）の技
術を基に拡張し、言語だけではなく時代や文
化をも横断する情報アクセス技術確立する
ことを目標として研究を行った。

本研究の成果は、主に人文学研究者による
利用を想定しており、人文学において利用
が進みつつある情報技術を用いて、人文学
研究を支援し、さらには人文学の研究手法
に新たな方向性を見出すことをもう一つの
目標とした。

3. 研究の方法

本研究では、研究対象の資料として、前
近代の日本語古典史料、伝統的モンゴル文
字で記述された史料、世界各地で公開され
ている浮世絵画像のデータベースにおける日
英のメタデータを対象として研究を行った。

日本語古典史料に対しては、現状では困
難であるテキストの単語への分割手法につ
いて研究を行った。また、古典史料から人
名や地名を抽出し、それらの出現頻度など
の統計的な情報を用いて人物関係を推定し
可視化する手法について研究を行った。

伝統的モンゴル文字文書に対しては、キ
リル文字による現代モンゴル語の問合せを
用いて、それとはまったく異なるスクリ
プトである伝統的モンゴル文字で書かれた
文書の検索を可能とする手法について研究
を行った。

浮世絵画像データベースに対しては、国
内で日本語のメタデータで公開されている
データベースのみならず、欧米の各国の美
術館・博物館などに所蔵されていて英語
のメタデータで公開されているデータベース
も対

象として、複数のデータベースに対して日
英二言語の双方から検索を可能とする手
法について研究を行った。

また、言語を跨った情報アクセスに繋
がる関連技術として、多言語 Web ページ
の作成支援の研究、今後増大する過去の
大量の情報から適切な情報を効率的に得
るための技術として、確率モデルに基づ
く情報推薦手法に関する研究を行った。

4. 研究成果

(1) 日本語古典史料のテキスト処理手法の 開発

日本語古典史料の自然言語処理のため
の基本技術の開発

古文の現代語による検索および古文テ
キストマイニングの実現に必要な現代語
- 古語間の対訳辞書を、古文とその現代
語訳の並列コーパスから自動的に構築す
る研究を行った。『源氏物語』の原文と
その現代語訳を用いた実験において、古
文テキスト中に現れる固有名詞を除去す
ることにより、除去しない場合と比較し
て精度が向上することが確かめられた。

また、言語資源に乏しい日本語古典史
料のテキスト処理の手法として、文字 N
グラム出現確率に基づく単語抽出手法
について研究を行い、名詞の抽出におい
て一定の精度を得ることができた。

古典史料に対するテキストマイニング
および可視化

主に日記形式の古記録から、人名・地名
などの固有表現を抽出し可視化する二
つの手法について研究を行った。

一つは、人名の共起関係を通じた人物
関係の推移の取得と可視化である。こ
れは、ある注目する二人の人物間の関
係の年ごとの推移を、別の第三者とこ
れら二人のそれぞれの共起関係から推
定する手法である。この手法では、同
じ人物の組に対して、同時代に書かれ
た複数の古典史料での可視化結果を比
較することも可能である。

もう一つは、人名・地名を通じた人物
関係の取得と可視化である。これは、あ
る人物に対して、その人名と共起した
各地名との共起頻度を次元とする人物
ベクトルを作成し、その類似度によっ
て人物関係の近さを推定する手法であ
る。さらに人物ベクトルをクラスタリ
ングすることで、より関係に近い人物
同士をグループ化することが可能であ
る。

これらの可視化手法を Web ブラウザ
上でインタラクティブに実行できる環
境を実装

した．このシステムを用いて人文系研究者による利用者評価実験を行った結果，概ね肯定的な評価を得ることができた．

(2) 人文系データベースの横断検索システム

美術館・博物館・大学などの各種機関で公開されている浮世絵の画像データベースを対象として，国内外の複数言語からなる異種データベースを横断検索する手法について研究を行った．

各機関において様々な異なるスキーマで提供されているデータベースを横断検索するために，メタデータの項目名の類似度を基にした自動マッピングの手法を開発した．また，日英の双方からの検索を実現するため，人名などを含む浮世絵画像のメタデータを翻訳する手法を実装した．

これらの手法を基に，実際に大英博物館，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，国立国会図書館などのデータベースを横断検索可能なプロトタイプシステムを完成させた．

(3) 伝統的モンゴル文字文書のデジタル図書館システム

キリル文字で記述される現代モンゴル語と異なり，伝統的モンゴル語は，縦書きでかつ行が左から右に進行し，単語内の位置によって文字の形が変化するなど，独自の書記体系を持っている．この伝統的モンゴル文字で記述された古典史料を保存するため，デジタル図書館システムを構築した．

このために，様々な文字コードやレンダリング手法が混在している伝統的モンゴル文字の現状において，互換性や将来的なテキスト処理のために最適な手法について調査し，最適な文字コードとそのレンダリング手法を確立した．

また，キリル文字による現代モンゴル語の問合せを用いて伝統的モンゴル文字文書を検索する，スクリプト横断型の検索手法を開発した．

これらの手法を実装し，インターネット上で公開した．また，モンゴル文字の研究者および一般利用者による評価実験を行った結果，システムの有用性および意義について非常に高い評価を得ることができた．

(4) 多言語 Web ページの作成支援システム

言語を跨った情報アクセスに繋がる関連技術として，Web サイトの多言語化を支援するため，文章の構文や意味の情報を保存することで機械翻訳の精度を向上させ，多言語

Web ページの作成を支援するシステムを開発した．本研究の成果について，国際会議 IMECS2012 において Best Student Paper Award を受賞した．

(5) 確率モデルに基づく情報推薦手法

今後増大する過去の大量の情報から適切な情報を効率的に得るための技術として，潜在的ディリクレ配分法 (LDA) に基づく評判情報のトピックモデリングおよび確率的潜在意味解析 (PLSI) を用いた文書分類およびその情報推薦手法への応用について研究を行った．

本研究の成果について，国際会議 IMECS2012 において Best Student Paper Award を受賞した．

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3件)

著者名：Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, Fuminori Kimura, and Akira Maeda，論文表題：A Study of Traditional Mongolian Script Encodings and Rendering: Use of Unicode in OpenType fonts，雑誌名：International Journal of Asian Language Processing，査読：有，巻：21(1)，発行年：2011，ページ：23-43

著者名：Akira Maeda and Yukinori Hayashi，論文表題：Classifying Web Content Using Discriminant Efficiencies，雑誌名：Journal of Information Technology Review，査読：有，巻：1(4)，発行年：2010，ページ：169-178

著者名：Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda，論文表題：Providing Universal Access to Japanese Humanities Databases: An Approach to Federated Searching System Using Automatic Metadata Mapping，雑誌名：Journal of Zhejiang University-SCIENCE C，査読：有，巻：11(11)，発行年：2010，ページ：837-843

[学会発表](計37件)

発表者名：Hiroshi Urae, Taro Tezuka, Fuminori Kimura, and Akira Maeda，発表標題：Structural and Semantic Indexing for Supporting Creation of Multilingual Web Pages，学会名等：International MultiConference of Engineers and Computer Scientists 2012 (IMECS2012)，発表年月日：2012年3月16日，発表場所：香港(中国)

発表者名：Takuya Konishi, Taro Tezuka, Fuminori Kimura, and Akira Maeda, 発表標題：Estimating Aspects in Online Reviews Using Topic Model with 2-Level Learning, 学会名等：International MultiConference of Engineers and Computer Scientists 2012 (IMECS2012), 発表年月日：2012年3月14日, 発表場所：香港(中国)

発表者名：吉村衛, 木村文則, 前田亮, 発表標題：古文テキスト解析のための文字Nグラムの出現確率を利用した単語分割, 学会名等：人文科学とコンピュータシンポジウム, 発表年月日：2011年12月11日, 発表場所：龍谷大学大宮キャンパス(京都府)

発表者名：久山岳夫, Biligsaikhan Batjargal, 木村文則, 前田亮, 発表標題：浮世絵を対象とした異種データベースの多言語統合アクセス手法の提案, 学会名等：人文科学とコンピュータシンポジウム, 発表年月日：2011年12月11日, 発表場所：龍谷大学大宮キャンパス(京都府)

発表者名：井坪将, 木村文則, 前田亮, 発表標題：古典史料からの相対的な人物関係の時間的変化の推定と可視化, 学会名等：人文科学とコンピュータシンポジウム, 発表年月日：2011年12月10日, 発表場所：龍谷大学大宮キャンパス(京都府)

発表者名：Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda, 発表標題：Metadata-related Challenges for Realizing Federated Searching System for Japanese Humanities Databases, 学会名等：11th International Conference on Dublin Core and Metadata Applications (DC-2011), 発表年月日：2011年9月22日, 発表場所：デン・ハーグ(オランダ)

発表者名：Sho Itsubo, Takahiko Osaki, Fuminori Kimura, Taro Tezuka, and Akira Maeda, 発表標題：Visualization of Co-occurrence Relationships Using the Historical Persons and Locational Names from Historical Documents, 学会名等：Digital Humanities 2011, 発表年月日：2011年6月20日, 発表場所：スタンフォード(アメリカ)

発表者名：Takaaki Shiida, Taro Tezuka, Fuminori Kimura, and Akira Maeda, 発表標題：Clustering of Review Comments for Restaurant Recommendation Using Probabilistic Latent Semantic Indexing, 学会名等：2011 International Conference on Data Engineering and Internet Technology (DEIT2011), 発表年月日：2011年3月17日, 発表場所：クタ(インドネシア)

発表者名：Takahiko Osaki, Sho Itsubo,

Fuminori Kimura, Taro Tezuka, and Akira Maeda, 発表標題：Visualization of Relationships among Historical Persons Using Locational Information, 学会名等：10th International Symposium on Web and Wireless Geographical Information Systems (W2GIS2011), 発表年月日：2011年3月3日, 発表場所：京都大学医学部芝蘭会館(京都府)

発表者名：Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda, 発表標題：A Survey on Rendering Traditional Mongolian Script, 学会名等：International Conference on Asian Language Processing (IALP2010), 発表年月日：2010年12月30日, 発表場所：ハルビン(中華人民共和国)

発表者名：井坪将, 大崎隆比古, 木村文則, 手塚太郎, 前田亮, 発表標題：古典史料における人名・地名情報を用いた共起関係の可視化, 学会名等：人文科学とコンピュータシンポジウム, 発表年月日：2010年12月12日, 発表場所：東京工業大学大岡山キャンパス(東京都)

発表者名：Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda, 発表標題：Approach to Cross-Language Retrieval for Japanese Traditional Fine Art: Ukiyo-e Database, 学会名等：14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2010), 発表年月日：2010年9月8日, 発表場所：グラスゴー(イギリス)

発表者名：Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, Fuminori Kimura, and Akira Maeda, 発表標題：An Approach to Ancient-to-modern and Cross-script Information Access for Traditional Mongolian Historical Collections, 学会名等：Digital Humanities 2010, 発表年月日：2010年7月9日, 発表場所：ロンドン(イギリス)

発表者名：Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, Fuminori Kimura, and Akira Maeda, 発表標題：An Ancient-to-modern Information Retrieval for Digital Collections of Traditional Mongolian Script, 学会名等：12th International Conference on Asia-Pacific Digital Libraries (ICADL2010), 発表年月日：2010年6月22日, 発表場所：ゴールドコースト(オーストラリア)

発表者名：Akira Maeda, 発表標題：Digital Humanities in the Context of Digital Libraries, 学会名等：Third International Conference on Digital Libraries (ICDL2010), 発表年月日：2010年2月24日, 発表場所：ニューデリー(インド)

発表者名：木村文則, 前田亮, 発表標題：固有名詞の検出による古文並列コーパスを利用した時代横断対訳辞書の構築, 学会名等：人文科学とコンピュータシンポジウム, 発表年月日:2009年12月19日, 発表場所：立命館大学びわこ・くさつキャンパス(滋賀県)

発表者名：井坪将, 木村文則, 手塚太郎, 前田亮, 発表標題：古典史料からの知識獲得および情報の可視化, 学会名等：人文科学とコンピュータシンポジウム, 2009年12月19日, 発表場所：立命館大学びわこ・くさつキャンパス(滋賀県)

発表者名：大崎隆比古, 木村文則, 手塚太郎, 前田亮, 発表標題：古記録の地名情報によるテキストマイニング手法の提案, 人文科学とコンピュータシンポジウム, 2009年12月19日, 発表場所：立命館大学びわこ・くさつキャンパス(滋賀県)

発表者名：Fuminori Kimura, Takushi Toba, Taro Tezuka, and Akira Maeda, 発表標題：Federated Searching System for Humanities Databases Using Automatic Metadata Mapping, 学会名等：9th International Conference on Dublin Core and Metadata Applications (DC-2009), 発表年月日:2009年10月14日, 発表場所：ソウル(韓国)

発表者名：Fuminori Kimura and Akira Maeda, 発表標題：An Approach to Information Access and Knowledge Discovery from Historical Documents, 学会名等：Digital Humanities 2009, 発表年月日:2009年6月23日, 発表場所：カレッジパーク(アメリカ)

〔図書〕(計3件)

著者名：前田亮, 木村文則, Batjargal Biligsaikhan, 出版社名：ナカニシヤ出版, 書名：デジタル・アーカイブの新展開(第9章：デジタル図書館・アーカイブへの言語・時代・文化横断型の情報アクセス), 発行年：2012, ページ：150-167

著者名：Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, Fuminori Kimura, and Akira Maeda, 出版社名：InTech, 書名：Digital Libraries - Methods and Applications (Chapter 2: Integrated Information Access Technology for Digital Libraries: Access across Languages, Periods, and Cultures), 発行年：2011, ページ：23-44

著者名：Fuminori Kimura, Akira Maeda, Kenji Hatano, Jun Miyazaki, and Shunsuke Uemura, 出版社名：Springer-Verlag, 書名：Trends in Communication Technologies and Engineering Science (Chapter 8: Utilizing Web Directories for Translation Disambiguation in Cross-Language

Information Retrieval), 発行年：2009, ページ：95-117

〔その他〕

ホームページ等

「伝統的モンゴル文字文書のデジタル図書館システム」

<http://tmsdl.media.ritsumeai.ac.jp/>

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