デジタル図書館・アーカイブ・ミュージアムのデータ共有化に関する研究

前田　亮（情報理工学部）

1. 研究目的

本研究プロジェクトでは、Web上に散在するデジタル図書館、デジタルアーカイブ、デジタルミュージアムなどに所蔵されている芸術・文化分野の各種デジタル資料を対象として、これらのデータの共有化の実現、さらにこれらに対する検索や各種分析を行うための情報技術に関する研究を行った。

具体的には、浮世絵の画像データベースを対象として、異なる言語およびメタデータスキーマからなる複数のデータベースを横断検索する手法について研究を進めた。また、これを発展させ、複数の浮世絵画像データベースの関連レコード間に自動でリンクを生成する手法、さらに複数データベース間における同作品を自動的に同定する手法について研究を行った。

また、日本語の古典史料テキストを対象として、これらのテキストから単語の区切りや人物情報を自動的に抽出するテキストマイニング手法、さらに人物と地名の関係性の推定に基づく人物関係の抽出および可視化手法について研究を行った。

また、日本だけではなく、モンゴルの歴史資料を研究の対象とし、伝統的モンゴル文字のコンピュータ上でのレンダリング手法の開発と、これを基にしたデジタル図書館システムの構築を行った。

また、インターネット上の各種メディア情報の共有化および芸術・文化分野の研究資源としての活用を目指した研究として、多言語Webページの作成を支援する手法、擬音語とリズム入力を用いた楽曲検索手法、電子掲示板の投稿内容から事物に対する比較評価表現を自動的に抽出する手法、プレゼンテーション資料中に含まれる図形が表す意味を推定することにより図形の検索を行う手法、画像集合に対して確率モデルに基づく特徴抽出を行い、画像の潜在的な意味構造を推定する手法、デジタルコンテンツに対するユーザエクスペリエンスを向上させる効果が期待されるゲーミフィケーション（非ゲームコンテンツのゲーム化）の手法などについて研究を行った。

2. 研究内容

2. 1　芸術・文化分野の各種デジタル資料の横断検索システム

本研究では、主に浮世絵の画像データベースを対象として、三つの研究を行った。

一つ目は、異なる言語およびメタデータスキーマからなる複数のデータベースに対して、項目名の文字列類似度の情報を用いることで、標準的なメタデータスキーマであるDublin Coreに自動的にマッピングする手法の開発と、それを利用した複数データベース間の横断検索手法の開発である。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、ボストン美術館を含む世界の全10機関がインターネット上で公開している浮世絵データベースと、国立国会図書館および米国議会図書館が所蔵している関連資料を含めた横断検索を行うプロトタイプシステムを作成した。

二つ目は、複数の浮世絵画像データベースから関連するレコードを自動的に見つけ出し、これらの間に自動でリンクを生成する手法であり、これは近年Linked Dataとして注目されている技術である。これを実現するために、各国の国立図書館等が提供している典拠データを用い、異なるデータベース間での同一作者による浮世絵間のリンクや、さらにその作者の関連情報へのリンク

の自動生成を実現した。本手法による人物同定の精度の定量的評価を行った結果、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。

三つ目は、複数の浮世絵データベースに存在する同一作品を自動的に同定する手法である。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。実際に浮世絵の作品名の音訳と英訳に対して提案手法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。

2.2 日本語古典史料からのテキストマイニングおよび可視化

本研究では、主に平安・鎌倉時代に書かれた日本語の古典史料の電子テキストを対象として、三つの研究を行った。

一つ目は、古典史料テキストから単語の区切りを自動的に見つけ出す手法である。テキスト処理において、単語の抽出は基本的な処理であるが、日本語は単語の区切りが明確でなく、現代日本語においては形態素解析の技術を用いることで単語を取り出すことが可能である。しかしながら、古い日本語については現代日本語用の形態素解析技術がそのままでは適用できず、単語を取り出すことは容易ではない。そこで本研究では、テキスト中の任意の長さの文字の並び（文字Nグラム）の出現確率の情報を用いて単語の区切りを推定する手法を開発した。

二つ目は、古典史料テキスト中から人物を表す表現を自動的に抽出する手法である。現代日本語に対して人物や地名などの固有表現を自動的に抽出する手法はすでに存在するが、古い日本語に対しては、上述の単語区切りの問題があり、有効な手法が開発されていないのが現状である。そこで本研究では、文字単位で固有表現抽出を行う既存の機械学習手法を基に、上述の単語区切りの推定結果の情報を組み合わせることで、古典史料テキストから人物を表す表現を自動的に抽出する手法を開発した。人物表現の自動抽出の精度について評価実験を行った結果、F値において最大で86.2%の精度が得られた。

三つ目は、人物と地名の関係性の推定に基づいて人物関係を自動的に抽出し可視化する手法である。具体的には、平安・鎌倉時代に書かれた古典史料である『兵範記』『吾妻鏡』『玉葉』を対象とし、これらの電子テキストおよび人名・地名索引の情報を用い、人名と地名が文中で共に現れた（共起）回数の統計情報を基にして、各人物を地名を次元とするベクトルで表現する。このベクトルが近い人物は、地名との共起の傾向が近いため、関係が近い人物であると推定できる。この手法を用いて人物間の関係を推定し、さらに、関係の近い人物をグループ化するクラスタリングの手法を組み合わせることで、人物間の関係を2次元のグラフとして可視化する手法を開発した。

2.3 伝統的モンゴル文字文書の文字レンダリング手法およびデジタル図書館の構築

本研究では、伝統的モンゴル文字で記述されたモンゴルの歴史資料を対象とし、これらの電子テキスト化に必要な文字符号系の検討および文字レンダリング手法の開発を行った。さらに、これらの文書に対する容易なアクセス手段を提供するために、キリル文字による現代モンゴル語のキーワードを用いて伝統的モンゴル文字で記述された文書を検索する手法を開発し、これを実装したデジタル図書館システムをインターネット上で公開した。

本研究で提案した伝統的モンゴル文字文書の現代モンゴル語による検索手法の検索性能について評価実験を行った結果、適合率で96.72%、再現率で77.85%と、実用的な性能が得られることを確認した。

2.4 各種メディア情報の共有化および芸術・文化分野の研究資源としての活用

本研究では、テキストだけではなく画像、楽曲、図形などの各種メディア情報を共有化し、さらに

は芸術・文化分野の研究資源としての活用を目指して研究を行った。

まず、芸術・文化分野の情報を世界に発信することを支援する技術として、多言語Webページの作成を支援する手法の開発を行った。本手法では、言語によって異なる構文や意味の情報を保存し、これらの情報を既存の機械翻訳技術と組み合わせることで、機械翻訳のみでは誤訳や不自然な訳となってしまう文に対しても、正しく自然な訳を生成することを可能にした。

また、楽曲情報を対象として、利用者による擬音語とリズム入力を用いた楽曲検索手法を開発した。本手法では、インターネット上での公開が進んでいる楽曲の演奏情報（MIDIデータ）を対象として、歌詞やアーティスト名などのメタデータによる検索だけではなく、楽曲のリズムやメロディの情報を用いて容易に楽曲検索を可能にするシステムを構築した。

また、インターネット上の電子掲示板の投稿から、複数の事物に対する比較評価表現を自動的に抽出する手法を開発した。電子掲示板上では、芸術・文化分野を含む様々な事物に対する批評や議論が行われているが、本研究では、これらの投稿から、複数の事物を比較して批評や感想などを述べている表現（対象・属性・評価表現の組）を自動的に抽出する手法を開発した。

さらに、写真や絵画などの画像データの集合に対して確率モデルに基づく特徴抽出を行い、画像の潜在的な意味構造を推定することにより、利用者の多様な意図に即した検索を可能とする画像検索手法や、芸術・文化分野を含むデジタルコンテンツに対するユーザエクスペリエンスを向上させる効果が期待されるゲーミフィケーション（非ゲームコンテンツのゲーム化）の手法について研究を行った。

3. 研究成果

芸術・文化分野の各種デジタル資料の横断検索システムの研究に関しては、国際会議11件、国内学会6件の発表を行った。また学術論文として、Literary and Linguistic Computing（Oxford Journals）を含む2件の論文が掲載された。なお、国際会議13th International Conference on Dublin Core and Metadata Applications（DC-2013）での発表に対して、Best Project Report Awardを受賞した。

日本語古典史料からのテキストマイニングおよび可視化の研究に関しては、国際会議8件、国内学会11件の発表を行った。また学術論文として、Literary and Linguistic Computing（Oxford Journals）に1件の論文が掲載された。なお、2013年の人文科学とコンピュータシンポジウムでのポスター発表に対して、Best Poster Award Bronze Prizeを受賞した。

伝統的モンゴル文字文書の文字レンダリング手法およびデジタル図書館の構築の研究に関しては、国際会議5件の発表を行った。また学術論文としてInternational Journal of Digital Library Systemsを含む2件が掲載された。

インターネット上の各種メディア情報の共有化および芸術・文化分野の研究資源としての活用を目指した研究に関しては、国際会議10件、国内学会22件の発表を行った。また学術論文として、情報処理学会論文誌：データベースを含む3件が掲載された。さらに、IAENG International Conference on Internet and Multimedia Technologies 2013での2件の発表に対して、それぞれBest Student Paper AwardおよびCertificate of Merit（Student）を受賞した。