様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 １４ 日現在

機関番号：１５５０１
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ００４６９

研究課題名（和文）翻刻・本文校訂を前提としないクラウド型古文書画像検索システムに関する研究

研究課題名（英文）A Study on Cloud-type Historical Document Image Retrieval System without Reprinting and Revision Process

研究代表者
中田　充（Nakata, Mitsuru）

山口大学・教育学部・教授

研究者番号：６０３０４４６６

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では、特徴グラフと呼ぶ文字構造情報に基づいて古文書画像を検索する技術の実現を目指して、（1）古文書画像を行単位に区切って特徴グラフに変換する技術、（2）部分グラフ同型判定問題を応用して、登録された大量の特徴グラフから、検索条件に類似した箇所を含むグラフを見つける技術、(3)その機能をインターネットを介して利用するためのWeb APIの設計ならびに試作版作成と、Web APIを利用したサンプルアプリケーションの実現を行った。

研究成果の学術的意義や社会的意義
古文書中の文字の形状を特徴グラフとして表現した上で、部分グラフ同定同形判定を活用して「似ているが少し違う文字を効率的に探し出す手法」を提案できたことには意味がある。この技術をより洗練した上で画像検索システムを実現すれば、テキストデータ化に膨大なコストをかけることなく、古文書画像検索が可能となり、これまで埋もれていた知識の発見に繋がる。
さらに、ここ2，3年注目されてきているAIを活用した画像検索技術と組合わせることで、高速かつ高い精度で「指定された文字と似ている文字を含む古文書」を検索可能となると期待できる。

研究成果の概要（英文）：In this research, the aim of our research is to realize an image retrieval system for Japanese historical documents which do not need reprinting and revision processes. We have developed the following technologies: (1) technology to divide a historical document image into each line, and convert it to feature graph; (2) technology for finding graphs that include similar structure to search conditions from many registered feature graphs by using the graph isomorphism determination problem; (3) Web APIs to generate and get feature graphs via the Internet. And we have realized a prototype version of document image register system using the Web APIs.

研究分野：情報工学，情報システム

キーワード：古文書画像検索　類似部分グラフ　同形部分グラフ　文字構造情報　特徴グラフ

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

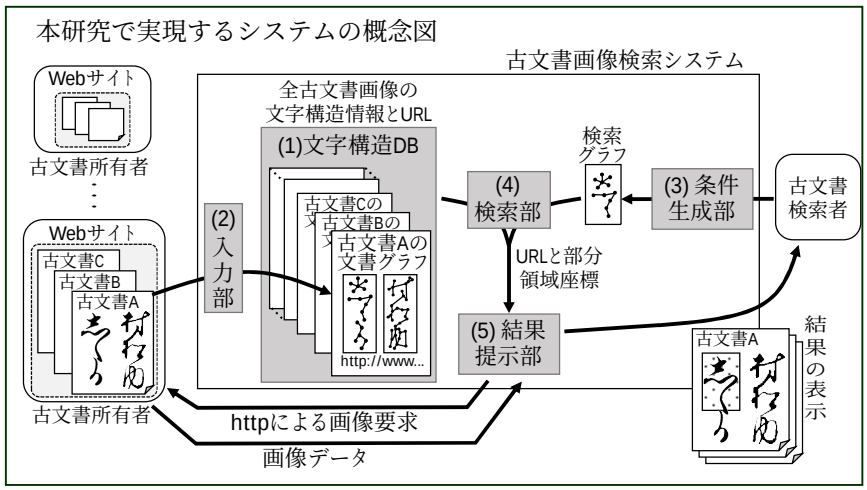
１．研究開始当初の背景

国や自治体等の博物館、文書館、図書館などが所蔵する様々な古文書の DB が構築され、インターネットを通じて広く一般に公開されている。しかし、全ての古文書が、翻刻（資料中の本文を現在の漢字や仮名に直す作業）や本文校訂（翻刻結果を補完・修正して本文を作り直す作業）を経た上で収録されている訳ではなく、DB 管理システムなどを用いずに、古文書の画像を資料名や項数に紐付けて Web 化しただけの DB が多数存在する。このような DB は、古文書の本文のテキストデータ（以降、本文テキスト）を含まないため、キーワード検索や全文検索などの検索機能を提供していない。そのため利便性が低く、有効に活用されているとは言い難い。

この原因として、翻刻・校訂作業には長い時間が必要であり、その従事者に高度な専門知識が求められるという問題点がある。そのため、和文古文書を対象とした文字認識技術を用いた翻刻支援の枠組みに関する研究がなされており、また申請者らも日本語続け字の認識技術を研究してきた。しかし、これらの認識精度は多種多様な古文書を自動認識するには不十分であり、翻刻者の負担を大幅に削減するには至っていない。さらに、利便性の高い DB の構築に必要な専門知識を持つ人員の確保が難しいという問題点もある。これらの問題は、予算や人的資源が限られている規模の小さい文書館などで顕著である。その結果、各地の文書館などが所蔵する「その価値が広く知られていない貴重な歴史資料」が参照されずに埋もれているという状況にある。このような背景の下、インターネット上に点在する、翻刻・本文校訂を行っていない古文書を簡単に登録し、その資料内容を検索可能とする技術が求められている。

２．研究の目的

本研究では、本文テキストの代わりに文字構造情報に基づいて古文書画像を検索する「古文書画像検索システム」を実現する（右図参照）。このシステムは、(1)古文書の文字構造情報の集まりである文字構造 DB、(2)古文書画像を文字構造情報に変換して DB に登録する入力部、(3)検索条件としての文字構造情報を生成する条件生成部、(4)文字構造情報に基づいた検索を実行する検索部、(5)検索結果の古文書画像を表示する結果提示部からなる。



文字構造情報とは 1 文字以上の文字列の形状を表す単純グラフである。文字構造 DB は古文書の行単位の文字構造情報（文書グラフ）と古文書画像の URL から構成される。入力部は古文書所蔵者が持つ古文書画像を行毎に分割した上で文書グラフに変換し、画像の URL と共に文字構造 DB に登録する。条件生成部は、検索者が所望する文字列の文字構造情報（検索グラフ）を生成する。検索部は、検索グラフに類似した構造を含む文書グラフを検索し、そのグラフに対応する古文書画像の URL と、画像に含まれる検索グラフに類似する部分領域の座標を求める。結果提示部は、URL を用いて古文書画像を取得し、検索条件に適合した部分領域を強調して表示する。

３．研究の方法

研究期間内に次の 4 つの研究課題に取り組む。なお、対象とする古文書は、続け字を含む毛筆の和文とし、各々の画像は既に Web ページ化されて URL を持つことを前提とする。

課題 1「古文書画像を文書グラフに変換する技術」：古文書画像を行単位に区切り、文書グラフに変換する技術を確立する。その際、行の傾きや重なり、汚れや裏写りに対応する補正技術も必要となる。この変換技術を任意の古文書画像の部分領域に適用することで、検索条件としての検索グラフを生成することも可能となる。

課題 2「検索グラフと類似する構造を含む文書グラフを検索する技術」：部分グラフ同型判定問題を応用して、登録された大量の文書グラフから、検索グラフと類似した箇所を含む文書グラフを見つける技術を確立する。本研究では、数行～三十行ほどで書かれた古文書の画像の集合から、数文字の文字列と類似した箇所を含む画像を見つけ出すことを想定している。したがって、サイズの大きい大量の文書グラフから、サイズの小さい検索グラフと形が似た部分グラフを余すことなく求める必要がある。さらに、手書きによる文字構造の微妙な違いに対応するために、「似ているが少し違う文字構造」を効率的に探し出す手法が必要となる。そのために、検索グラフを「条件として欠くことが出来ない構造を表す部分」と「それ以外の構造を表す部分」とに分けて、前者と同型の部分グラフを文書グラフから求める手法を提案する。

課題 3「文書グラフの索引技術」：古文書中の文字は行単位でグラフ化されるため、文字構造 DB 中の文書グラフの数は莫大になる。そのため、文書グラフを高速に検索するための索引技術が必要となる。本研究では、グラフの節点の位置や辺の方向などを考慮した索引・検索技術を提案する。

課題 4「インターネットを介して文書グラフと画像 URL を作成・登録・検索する Web API」：インターネットを介して、古文書所蔵者が所有する古文書画像から文書グラフを生成し、文書グラフと画像の URL を文字構造 DB に登録可能な技術を実現する。さらに、これらの技術を Web API として実現し、クラウド型古文書画像検索システムの実現を可能とする。

４．研究成果

古文書画像を文書グラフに変換する技術については、まず、古文書を行単位に分割する作業を半自動で行えるような支援ソフトウェアを作成した。このソフトウェアにより、裏写りのない三行以上に跨がった極端に幅の広い行を含まない画像についてはほぼ問題なく一行単位に分割できるようになった。次に、行単位に分割された画像から角度距離グラフを用いて行毎に文字の形状を表現する文書グラフを作成する技術を提案し実装した。

2 番の課題として挙げた検索グラフと類似する構造を含む文書グラフを検索する技術については、検索結果に必ず含まれなければならない文字構造を表現した「必須グラフ」と検索条件を大まかに表現した「検索グラフ」の二種類の特徴グラフを導入し、文書グラフ中に含まれる「必須グラフと同型の部分グラフ」を見つけ出すことで、検索条件として指定された文字の形（構造）と「似ているが少し違う形をした部分画像」を検索する手法を提案し、プログラミング言語 Java を用いてその実装を行った。また、文書グラフの索引・検索技術については、特徴グラフの正規化を行い、その上で検索時間を軽減するためのアルゴリズムを提案した。

インターネットを介して文書グラフと画像 URL を作成・登録・検索する Web API については、プログラミング言語 PHP を用いて特徴グラフを生成・登録・検索する機能を Web API として実装した。この API は Web ページとして行単位で公開されている古文書画像の URL を引数として、文書グラフを生成・登録する機能と、指定された識別子を持つ文書グラフを原画像とともに表示する機能を持つ。さらに、これらの API を利用して画像登録機能を JavaScript で実現した。

今後は、検索グラフに類似した部分グラフを検索する技術の速度と精度の向上を図り、最近注目を浴びている AI を活用した画像検索と共に活用することで、本研究の背景で挙げた「翻刻・本文校訂を前提としない古文書検索技術」の実現に繋がる。

５．主な発表論文等

〔雑誌論文〕（計 4 件）

1. 肥喜里 大地、中田 充、葛 崎偉、吉村 誠、同型部分グラフの判定に基づいた古文書画像切り出し技術の提案、第 31 回 回路とシステムワークショップ論文集、査読有り、2018、pp.299-304
2. Hiroaki Kodama, Mitsuru Nakata, Qi-Wei Ge, and Makoto Yoshimura、Improvement of Generation Method of Feature Graph Representing Japanese Handwritten Character String、Proc. of ITC-CSCC2017、査読有り、2017、pp.243-246
3. Hiroaki Kodama, Mitsuru Nakata, Qi-Wei Ge, Makoto Yoshimura、Implementation and Evaluation of Similar Subgraph Retrieving、Proc. of ITC-CSCC2016、査読有り、 2016、pp.281-284
4. Hiroaki Nagaoka, Mitsuru Nakata, Qi-Wei Ge and Makoto Yoshimura 、Similar Subgraph Retrieving for Japanese Historical Document Search System、Proc. of ITC-CSCC2015、査読有り、2015、pp.222-225

〔学会発表〕（計 2 件）

1. 児玉 啓彰、中田 充、古文書画像検索システムのための類似部分グラフ検索手法の改良、電子情報通信学会システム数理と応用研究会、2017
2. 児玉 啓彰, 中田 充、古文書画像検索システムのための類似部分グラフ検索法の実現、電子情報通信学会システム数理と応用研究会、2016

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：葛　崎偉
ローマ字氏名：Qi-Wei Ge
所属研究機関名：山口大学

部局名：教育学部
職名：教授
研究者番号（8桁）：30263750

研究分担者氏名：吉村　誠
ローマ字氏名：Yoshimura Makoto
所属研究機関名：山口大学
部局名：教育学部
職名：教授
研究者番号（8桁）：70141116

(2)研究協力者
研究協力者氏名：肥喜里　大地
ローマ字氏名：Hikiri Daichi

研究協力者氏名：児玉　啓彰
ローマ字氏名：Kodama Hiroaki

研究協力者氏名：鷹多　穂実
ローマ字氏名：Takata Honomi

研究協力者氏名：山本　怜子
ローマ字氏名：Yamamoto Reiko

研究協力者氏名：長岡　弘祥
ローマ字氏名：Nagaoka Hiroyuki