様式Ｃ－１９

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　６月　１日現在

機関番号：３２６０５
研究種目：若手研究（B）
研究期間：２０１０～２０１１
課題番号：２２７２０２３９
研究課題名（和文）　利用者本位の古文書デジタルアーカイブを実現する情報検索技術

研究課題名（英文）　User friendly information retrieval technologies for historical document digital archives

研究代表者
未代　誠仁（KITADAI AKIHITO）
桜美林大学・総合科学系・講師
研究者番号：００４０１４５６

研究成果の概要（和文）：古代木簡デジタルアーカイブをターゲットとした知的情報検索の実現に向けた研究を行った。パターン認識技術、画像処理技術、ペン・ユーザインタフェース技術を用いた破損字形検索技術を実現することで、解読が困難となった字形の類例検索を実現した。また、検索技術を含む古代木簡解読支援システム「Mokkanshop」の開発と一般公開を行った。このソフトウェアは古代木簡デジタルアーカイブ「木簡字典」と連動し、利用者に実用的な情報検索を提供する。

研究成果の概要（英文）：
This research was to provide intelligent information retrieval for historical mokkan digital archives. We had implemented a character pattern retrieval technology for damaged and unreadable character patterns by combining the technologies of pattern recognition, image processing and pen-based user interface. Also, we had developed and provided software “Mokkanshop” to support reading the historical mokkans. This software is now working with our digital archive of the historical mokkan to provide useful information for the users.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,800,000	540,000	2,340,000
2011 年度	1,400,000	420,000	1,820,000
年度			
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：情報考古学、パターン認識、ヒューマンインタフェース
科研費の分科・細目：史学・日本史
キーワード：情報検索、デジタルアーカイブ

１．研究開始当初の背景

古文書デジタルアーカイブには２つの大きな役割がある。1 つは、現在進行形で劣化し続ける古文書の今ある姿を後世に残すこと（文化的価値の保存）であり、もう１つは古文書が持つ史料としての価値を高めること（史料的価値の向上）である。しかし、デジタルアーカイブの収録点数が増加すると、文化的価値が高まる反面、必要な古文書を探し出すのが困難となり史料的価値は低下する。この問題を解決するには収録点数に応じた情報検索技術の高度化が不可欠であった。

研究代表者は、奈良文化財研究所の史学者らと共に古代木簡デジタルアーカイブ「木簡字典」の構築を進めてきた。8 世紀前後に作成・使用された古代木簡は、平城宮跡だけで約 18 万点、日本全体では約 32 万点が発見されており、奈良時代前後の言語/地名/人名/産業/地域間交流などの検証、および同時代に記述された様々な古文書（難読な古代木簡を含む）の解読を行う上で極めて有用な参考史料となっていた。

木簡字典の収録点数は順調に増加を続けてきた。研究代表者も、文字認識/画像処理/釈文補完などの機能を搭載した古代木簡解読支援システムの開発を通して古代木簡が持つ文化的価値の保存に微力を尽くしてきた。同時に、木簡字典を Web 上で一般公開し、利用者からの意見を広く集めながら、木簡字典/古文書デジタルアーカイブの史料的価値を向上させる技術の検討を進めてきた。その中で、人の曖昧な記憶/劣化が著しい古文書に含まれる僅かな情報（断片的な釈文/不完全な字体など）からでも必要な古代木簡に辿り着ける知的情報検索技術の実現が重要であるとの結論に達したのだった。

古文書デジタルアーカイビングに関するこれまでの研究は文化的価値の保存に重きを置いたものであった。それらが一定の成果を上げつつある中で、効果的な情報検索の実現は必然ともいえる課題になっていた。古代木簡デジタルアーカイブに対しては、史学者/考古学者のみならず、書家/歴史愛好家といった多くの人々が興味を持つに至っていた。このような時期にこそ、古文書デジタルアーカイブの研究に関わるすべての者が利用者本位の見地に立ち、史学発スマート・ユビキタスネットワークの実現を目指すべきと研究代表者らは考えていた。そこで、研究代表者らはキー文字列の誤字/脱字/語順変動に柔軟に対応できる曖昧テキスト検索のための Extended Aho-Corrasick 法（EAC 法）、不完全な字体をキーとする破損字体検索のためのテンプレート修正法などを実現し、利用者が持つ僅かな情報からでも必要な古文書を探し出せる知的情報検索技術の実現を目指してきた。これらの手法は、8 世紀前後の古文書を用いたベンチマークにおいて高い効果を示すに至った。

２．研究の目的

古文書デジタルアーカイブの価値を決定付けるのは、充実したコンテンツと、それに応じた高度な検索技術である。

本研究では人の曖昧な記憶、劣化が著しい古文書などに含まれる僅かな情報から必要な古代木簡を探し出す「知的情報検索技術」の実現を通して、古代木簡デジタルアーカイブにおける文化的価値の保存と史料的価値の向上が両立可能であることを示すことを主たる目的の一つとした。

また、適切な情報検索技術と対話的なユーザインタフェースを組み合わせることで、人の知性に直結した利用者本位のデジタルアーカイブが実現可能であることを明らかにすることとした。

さらに、これらの目標を達成することで、古文書デジタルアーカイブが古代文明に対する人々の理解を深め、興味を喚起し、文化の継承に資する存在であることを示すことを意義的な目的とした。

３．研究の方法

本研究については、(1)文脈処理/破損字体検索の高精度化、(2)思考的負担を軽減する対話的なユーザインタフェースの構築、(3)ノイズを含むキーに対する情報洗浄技術の実現、(4)木簡字典（古代木簡デジタルアーカイブ）の仕様拡張/メタデータ作成、の 4 項目、および利用者を対象とした評価実験から構成することとした。実施に際しては、奈良文化財研究所史料研究室、東京農工大学中川正樹研究室、国際日本文化研究センター山田奨治研究室、東京海洋大学古谷雅理助教らの協力を得ることとした。また、本研究では研究代表者が研究分担者となっている別の研究課題（古代木簡デジタルアーカイブ/解読支援の実現に関する研究）との間に相乗効果が得られるよう工夫しながら、古文書解読支援に限定しない広い意味での史料的価値向上を図り、古代文明に対する一般人の理解促進/将来に渡る日本文化の継承に資することのできる、多くの人にとって有益な情報検索技術の実現を目指すこととした。

４．研究成果

2010 年度は、字体検索の高度化を達成すべく、墨/背景分離のための画像処理、および絞り込み検索との親和性が高い字体認識技術を中心とした研究ならびに成果発表を行った。画像処理技術は前章の(1)に対する中核技術となった。また、字体認識に対する研究は(3)の破損字体検索の高精度化に貢献すると共に、その結果を入力とする文脈処理の高度化にも寄与した。

画像処理では色空間の変換、周波数帯分離など様々な手法の効果を精査し、その成果を国内外の学会において発表した。また、字体認識技術では筆の太さの変化による誤認識に頑健な手法を開発し、利用者が筆跡を補完・推定しながら字体検索を行う際の精度向上を達成した。

画像処理では、墨の退色と木片（背景）の変色により形状が曖昧になった字体を先鋭化しつつ、同時に字体以外の部分にノイズが乗るのを防ぐことが重要である。これらは基

本的に相反する要素であるが、本研究では特性の異なる複数の画像処理を併用することで両立できることを明らかにした。(ただし、実際の運用においてユーザに複数の画像処理を選択実行させるのはユーザビリティの面から好ましくなく、この点については2011年度の研究目標とした)。

字体認識については画像処理との有機的な連携を通して精度を高められるよう改良を行った（この成果を活用し、絞り込み検索におけるユーザビリティを画像処理と共通化することで、字体の先鋭化と検索をシームレスに実施できるユーザインタフェースを実現することについては 2011 年度の目標とした)。

2011年度は情報検索精度の向上、検索対象の拡大、画像処理の改善、および使い勝手(ユーザインタフェース）の改善を行った。これは前章の(1)、(2)、(4)に対する中核技術となった。

検索精度の向上については、破損を伴う字形の評価基準として墨濃度の勾配特徴を導入し、字形輪郭部の荒れ、にじみなどに頑健な類似度計算技術を実現した。勾配特徴は字形筆記時の運筆方向の推定に確率的なソフトディシジョンの導入を可能にする。ここで必要となる確率モデルの構築では、実際に古代木簡から得られた字形情報（後述）を活用した。また、評価結果として得られる類似度（確率値）に経験知に基づく後処理を加えることで文脈処理とのスムーズな連携に考慮した。

検索対象の拡大では、古代木簡データベース『木簡字典』の拡張によって得られた多数の古代木簡画像から新たな字形情報を抽出・整理することで実現した。この成果は検索対象となる古代木簡の拡大だけでなく字形評価技術の精度向上にも大きく貢献した（前述)。

画像処理の改善では、古代木簡の画像に対する分析結果を基に字形（墨）を木簡表面の木目、腐食、荒れなどと効果的に分離可能な処理技術を実現した。この画像処理技術では調整の必要なパラメータを最小限に留めたことが特徴である。これに伴い、今までは複雑なパラメータ調整が必要となっていた画像処理用インタフェースを全面的に置き換え、シンプルな作業で字形（墨）を抽出できるようにした。

これらの研究成果は、次章にある通り国内外の査読付論文誌、および学会発表などの場を通して広く認知されるに至っている。文書は人間にとって不可欠な存在であり、そのデジタルアーカイブは今後も規模、点数共に拡大を続けて行くであろう。その中で、本研究の成果は幅広い分野・領域に波及効果を及ぼすものであると考えている。

今後の展望については、本研究によって実現した技術をより多くの人に活用して貰うためのインフラストラクチャーの構築が急務となると考えている。コンピュータの形状、用法、およびそれを取り巻く環境はこの２年だけでも大きく変化した。これまで以上に多くの人が古文書デジタルアーカイブに触れ、その価値を感じることができる情報技術の実現に向けて研究活動を続けたいと考えている。

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　7件）

① Akihito Kitadai, Masaki Nakagawa, Hajime Baba and Akihiro Watanabe, Similarity Evaluation and Shape Feature Extraction for Character Pattern Retrieval to Support Reading Historical Documents, Proc. 10th IAPR International Workshop on Document Analysis and Systems-2012 (IEEE Computer Society), 査読有, 2012, pp.359-363

② 耒代 誠仁、中川 正樹、馬場 基、渡辺 晃宏、古代木簡解読支援のための画像処理および字体検索の高度化、情報処理学会人文科学とコンピュータシンポジウム「じんもんこん 2011」論文集、査読有、Vol.2011、No.8、2011、pp.93-98

③ 耒代 誠仁、木簡解読支援のための情報技術の研究を通して、木簡学会『木簡研究』、査読有、33巻、2011、PP.167-173

④ Sherini Somayeh、耒代 誠仁、中川 正樹、馬場 基、渡辺晃宏、古代木簡解読支援システムにおける字体検索の高性能化、人文科学とコンピュータシンポジウム論文集、査読有、Vol.2010、No.15、2010、pp.27-32

⑤ Jun TAKAKURA, Akihito KITADAI, Masaki NAKAGAWA, Hajime BABA and Akihiro WATANABE, Techniques to Enhance Images for Mokkan Interpretation, Proc. 12th International Conference on Frontiers in Handwriting Recognition, 査読有, Vol.1, No.1, 2010, pp.358-362

⑥ Akihito KITADAI, Sherini SOMAYEH, Masaki NAKAGAWA, Hajime BABA and Akihiro WATANABE, Non-linear Normalization of Damaged Characters for Search Refinement, Proc. 2nd China-Japan-Korea Joint Workshop on Pattern Recognition, 査読有, Vol.1, 2010, pp.185-188

⑦ Jun TAKAKURA, Akihito KITADAI, Masaki NAKAGAWA, Hajime BABA and Akihiro WATANABE, Techniques of Image Processing for Decoding Mokkans, Proc. 2nd China-Japan-Korea Joint Workshop on Pattern Recognition, 査読有, Vol.1, 2010, pp.47-49

〔学会発表〕(計 4件)

① Akihito Kitadai, Similarity Evaluation and Shape Feature Extraction for Character Pattern Retrieval to Support Reading Historical Documents, 10th IAPR International Workshop on Document Analysis and Systems-2012, 2012/03/29, Gold Coast, Australia

② 耒代 誠仁、古代木簡解読支援のための画像処理および字体検索の高度化、情報処理学会人文科学とコンピュータシンポジウム「じんもんこん 2011」、2011/12/10、龍谷大学 大宮キャンパス（京都）

③ 耒代 誠仁、木簡解読支援のための情報処理技術、木簡学会（招待講演)、2010年 12 月 4 日、奈良文化財研究所平城宮資料館講堂

④ Akihito KITADAI, Non-linear Normalization of Damaged Characters for Search Refinement, 2nd China-Japan-Korea Joint Workshop on Pattern Recognition, 2010 年 11 月 4 日, Fukuoka, Japan

〔その他〕
ホームページ等
奈良文化財研究所「木簡ひろば」
http://hiroba.nabunken.go.jp/index.html
（開発ソフトウェアの公開）

６．研究組織
(1)研究代表者
耒代 誠仁 (KITADAI AKIHITO)
桜美林大学・総合科学系・講師
研究者番号：00401456