研 究 報 告 書

「擬似コード変換と統計解析による文書画像からの知識抽出」

研究期間: 平成 20 年 10 月～平成 24 年 3 月
研 究 者: 寺沢 憲吾

1，研究のねらい

ディジタル技術の普及と発展に伴い、多くの文献・史料がディジタル画像化され公開されている。これにより従来は図書館や博物館の奥深くに所蔵されていた貴重な文献・史料を、一般のユーザがネットワークを介して容易に入手・閲覧することが可能となってきている。

このようにディジタル技術が文献・史料という知識の共有・流通に対して大きな成果を上げている一方で、これらのディジタル化された文書画像を活用するための情報処理技術は未だ発展途上段階である。文書画像に対して全文検索やテキスト解析・マイニングといった情報技術を適用するには、画像を正確にテキストデータに変換することが望ましいが、OCR（光学文字認識）技術を用いて文書画像をテキストデータ化する方法は歴史的文書や手書き文書を対象とする場合に常に適用可能であるとは限らない。OCR は言語や書体に依存した技術であるため、時代が古いものや特殊な文字あるいは用法を含むもの、また十分なサンプルが確保できないものには適用できない。また手書き文字の場合は精度が低下するため正確なテキストデータの作成のためには専門家の手による修正作業が不可欠であり、あらゆる文書画像をこの方法でテキストデータ化するのはコストの問題で現実的ではない。

本研究のテーマは、画像をテキスト化してから解析するのではなく、画像を画像のままで解析する手法を確立することである。具体的には、文書画像を画像特徴量による擬似コード表現に変換することにより、文書画像データに対する高速な全文検索法を開発するとともに、統計解析による知識抽出のための技術として、頻出語句の抽出、語句の頻度分析による文書の特徴づけ、特徴に基づく文書間の関連性の記述、共起関係等による語句間の関連性の記述などの方法を開発する。こうした一連の手法を確立させることにより、従来「取り扱いにくい」データであった画像データを、「取り扱いやすい」テキストデータと同様に知識創出のための情報資源として役立てることが可能となる。いわば、画像データとテキストデータとの間の架け橋である。この取り組みを通して、ディジタル文書画像という知識の宝庫を、共有・流通という第一段階から、それを基礎とした知識の抽出、さらには知識の創出という次の段階へ進めるということが本研究のねらいである。

2，研究成果

本研究は大きく分けて、文書画像を擬似コードに変換すること、擬似コードを用いた文書解析アルゴリズムを構築すること、それらを用いて実際に利用可能なアプリケーションを構築し、文献研究などの用途に役立てることの3つの柱からなる。以下にそれぞれの成果について述べる。

（1） 文書画像の擬似コードへの変換

本研究の核となる擬似コード LSPC は、文字の形を記述する画像特徴量（高次元の実数ベクトル）を、その記述性能を大きく損なうことなく、比較的低次元の自然数の組として表現する手法で

ある。この変換には、近傍探索問題の解法の1つである LSH のインデックスを用いるが、ここで、ノルムが1に正規化されているベクトルの集合に対して従来の LSH より有効な SLSH（Spherical LSH）を用いることで、擬似コードの性能をさらに高めている。この手法自体はさきがけ研究以前に開発したものだが、さきがけ研究期間中に、SLSH を実際に大規模データに対して適用した場合についての検証評価を行い、有用性を確認した。この研究は論文「Approximate Nearest Neighbor Search for a Dataset of Normalized Vectors」として発表し、電子情報通信学会論文賞を受賞し、高く評価された。

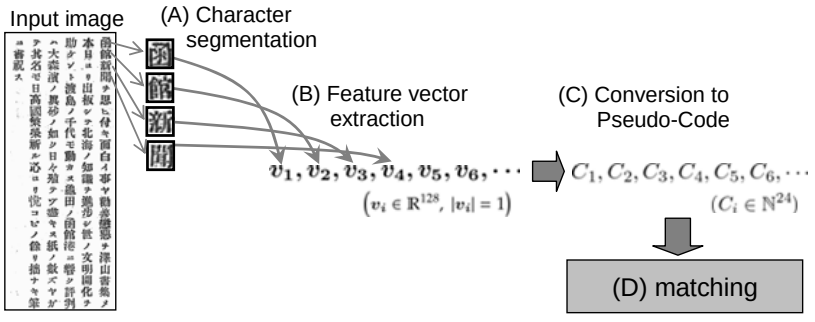
（2） 擬似コードを用いた高速検索手法の確立

擬似コードを用いることで、通常の画像ベクトルに基づく検索よりもきわめて高速な検索が可能であることを示した。「函館新聞」（北海道で最初の民間新聞・明治 11 年～明治 31 年発行）のうち明治 11～17 年の 7 年分、859 万文字から 5～10 文字の文字列を検索するのに、所要時間が1秒未満（線形探索を用いた場合）という高速検索を可能にした。また、Ex-BMH 法というアルゴリズムを開発し、文字列長が 9 文字以上の場合に、線形探索よりも高速な検索が可能であることを示した（図 1，図2，表1）。

（図1）「函館新聞」明治 14 年

（表1）「函館新聞」に対する全文検索時間

検索文字列長	線形探索[秒]	Ex-BMH[秒]
5 文字	0.90	1.23
6 文字	0.81	1.08
7 文字	0.90	1.01
9 文字	0.91	0.86
10 文字	0.87	0.81



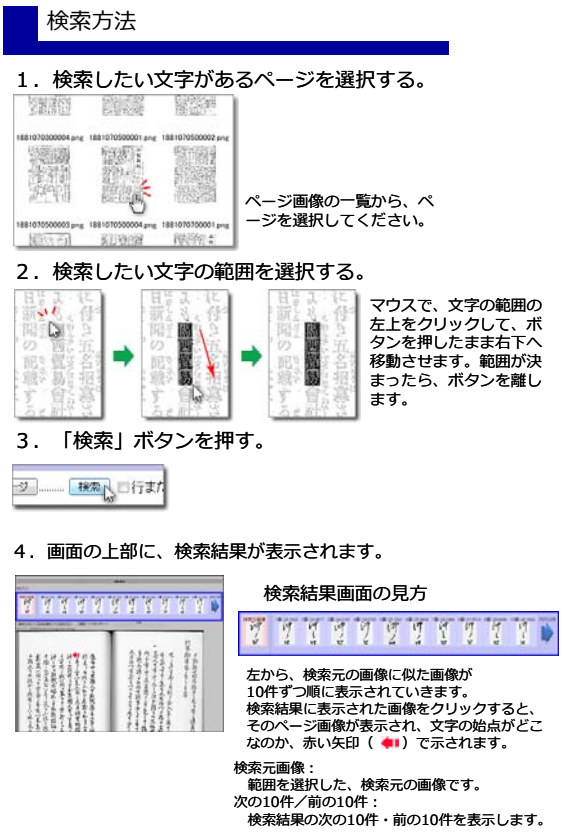
（図2）開発した手法の概略図

（3） 実用アプリケーションの構築

本研究で開発した全文検索システムをインターネットを通して一般のユーザが使用できるよう、ウェブシステムを構築し、平成 23 年 5 月に第1段階として、函館市中央図書館の協力を得て、同図書館所蔵の文献の中から、「亜国来使記」（1854 年 4 月 ペリーが箱館訪問した際の松前藩の応接記録）および「函館新聞」（北海道で最初の民間新聞・明治 11 年～明治 31 年発行）のうち明治 14 年の1年分を公開した。html サーバは専用のものを用意し、本学に設置した（http://records.c.fun.ac.jp/）（図3，図4）。また、函館市中央図書館デジタル資料館のページ（http://lib-hkd.jp/rein/）からもリンクを設定し、誘導した。さらに、平成 24 年 1 月には、京都大学文学研究科と連携し、京都学派アーカイブ（http://kyoto-gakuha.info/）で公開されているコンテンツの一部に対し、全文検索を可能にして公開した。

実用アプリケーションへ検索エンジンを提供する活動は上記以外にも複数行っている。一例として、京都大学文学部の林晋教授が開発している歴史学、文献学などの人文学におけるテキスト研究用のツールである SMART-GS（http://sourceforge.jp/projects/smart-gs/）へ検索エンジンを提供している。SMART-GS は検索の他にリンクやマークアップ、コメントの添付などの機能を持つ多機能なツールであるが、提供した検索エンジンは独立した部品として動作するよう意識して設計したため、SMART-GS の各機能の開発と検索エンジンの開発とは独立に行うことが可能であった。また別の例として、民間企業から産学連携の打診があり、この検索システムを活用した新しいウェブサービスの開発に着手している。

（図3）実際に公開されている「文書画像検索システム」のタイトル画面.
URL: http://records.c.fun.ac.jp/



（図4）オンライン検索システムの操作方法

3，今後の展開

擬似コードに対する解析アルゴリズムの開発に関しては、依然研究開発の余地が残っている。研究のために必要なデータや、基礎的な知見は既に得られているので、今後はこれをさらに進展させ、まだ未開発の部分の開発を進めていきたい。擬似コードに対して解析を行うことで文書の解析がある程度まで可能であることを自らの手で立証し、こうした研究に対するフォロワーを生み、研究手法の一つの潮流を確立させることが将来的な大目標である。

これまでに引き続き、人文系研究者と連携して、人文学の研究のために情報技術が提供できることを模索していく。情報システムのアシスト無しには見いだせなかった人文学上の新たな知見を継続して生み出せるようになれば大きな収穫である。

検索エンジンをウェブサービスとして公開させたので、すでに利用者からの意見や要望が寄せられつつある。今後はこうしたフィードバックをふまえ、システムの改良を図っていきたい。それとともに、コンテンツを増強させて研究成果のアピール力を高め、ディジタルライブラリーの利活用手法の一つとして定着するところまでを目指していきたい。

研究期間中に、産学連携で民間企業と共同研究を行うための交渉を進めることができたので、今後もこの活動を継続させていく。研究者自身の専門である画像処理とアルゴリズムの研究にプラスして、連携先企業の得意とするウェブサービスを融合させることで、新しい情報技術・ウェブサービスを実際に稼働させ、新たな知の創生の苗床とすることを目指していく。

4，自己評価

当初計画と比べ、研究成果を社会に還元する、研究のアウトリーチ活動は想定以上の進展を得た。これには、イノベーション・ジャパン 2009 での展示発表において多くの企業の方に本研究内容を紹介することができ、ウェブサービスを得意とする企業とのマッチングが生じ、産学連携を見据えた協力関係を構築することができたことが大きく貢献している。また、研究中に研究者自身が公立はこだて未来大学へ着任した後、函館市中央図書館との連携が円滑に進んだことも要因の一つである。さらには、研究領域アドバイザーの支援の下、人文学の研究者との協力関係を構築することもできた。また、当初計画より早期にウェブサービスを公開することができたため、一般ユーザの利用履歴データの解析にも着手することができた。

一方で、アルゴリズムの開発を中心とする理論的研究に関しては、部分的には成果が得られたものの、当初目標として掲げた計画を期間内にすべて達成するには至らなかった。ただし、この研究活動は今後も継続していく予定であり、今後当初計画にある成果を得るための足がかりとなる部分は確立できたのではないかと考えている。

そうした点を総括すると、この研究のねらいとして掲げた「画像データとテキストデータの間に橋を架けること」に関しては、研究期間内に、橋を架けることはできたと評価してよいのではないかと考えている。橋の上の交通量をさらに増やしていくための研究活動は今後も引き続き継続していく予定である。

5，研究総括の見解

文字認識不能な文書画像を画像のまま部分検索などできるようにする重要な基盤技術の提案である。実績は高く、技術としての有用性が高い課題であった。

研究成果を社会に還元する、研究のアウトリーチ活動は当初計画を超えた進展を得ている。函館市中央図書館との連携、人文学の研究者との協力関係を構築できている。さらにウェブサービスの公開により、一般ユーザの利用にもつながっていることは高く評価できる。情報検索の分野に大きく貢献したと考える。

この研究のねらいの「画像データとテキストデータの間に橋を架けること」に関しては達成している。これまでに存在しなかった研究分野を開拓した功績は大きい。人文科学における新しい道具ができたので、それを活用した研究成果も期待できる。今後、更に効率の良いアルゴリズムの開発を中心とする理論的研究を含め、この活動が個人を超えて発展することを期待する。

6, 主な研究成果リスト

（1）論文（原著論文）発表

1. K. Terasawa and Y. Tanaka, "Approximate Nearest Neighbor Search for a Dataset of Normalized Vectors," IEICE Transactions on Information and Systems, vol.E92-D, no.9, pp.1609-1619, 2009. （平成 21 年度（第 66 回）電子情報通信学会論文賞受賞）

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物等）

【学会発表】
1. K. Terasawa, T. Kawashima, Y. Tanaka, "The Extended Boyer-Moore-Horspool Algorithm for Locality-Sensitive Pseudo-Code," VISAPP 2011, International Conference on Computer Vision Theory and Applications (Part of VISIGRAPP 2011), pp.437-441, Algarve, Portugal, Mar. 5-7, 2011.
2. K. Terasawa, T. Shima, T. Kawashima, "A Fast Appearance-Based Full-Text Search Method for Historical Newspaper Images," ICDAR2011, 11th International Conference on Document Analysis and Recognition, pp.1379-1383, Beijing, China, Sept. 18-21, 2011.
3. 寺沢憲吾, 川嶋稔夫, "文書画像からの全文検索のオンラインサービス", 人文科学とコンピュータシンポジウム「じんもんこん 2011」, pp.329-334, 京都, 2011 年 12 月.
【公開したシステム】
1. 文書画像検索システム http://records.c.fun.ac.jp/