

画像データの分析から 歴史を探る

－「武鑑全集」における「差読」の可能性－

北本朝展（ROIS-DS 人文学オープンデータ共同利用センター／国立情報学研究所）

1. 文字と非文字

画像データとは、画素（ピクセル）が平面的または空間的に並んだデータです。しかしそれぞれの画素はその明るさを表す数値データを格納しているだけであり、その数値データの「意味」を取り出すことは簡単ではありません。人間は画像から意味を読み取る能力に長けているため、両者の違いを意識することはあまりないかもしれませんが、機械にとっては数値の世界と意味の世界とのギャップ（セマンティックギャップ）を越えることは今も難しい問題です。とはいえ、人間が画像から意味を読み取るという、昔からの画像データ利用を続けるだけでは、画像データというメディアの特性を十分に引き出すことはできません。人文学では、文字情報の有用性があまりに高いため、文字にばかり意識が向かってしまい、非文字情報への注目が不十分となりがちです。

非文字情報に注目した研究の例として、われわれが地図や写真を研究してきた経験から提案した考え方「デジタル史料批判（digitally-enabled criticism）」を紹介したいと思います。史料批判とは、史料の信頼性を批判的に検証する研究プロセスであり、史料がどのような由来を持ち、どのような表現が使われているかなどの点を、史料を相互比較しながら検証することで、史料がどのくらい信頼できるかを調べていきます。しかし「書かれたもの」は必ずし

も文字だけではありません。例えば地図はどうでしょうか。われわれは地図を相互比較する方法で地図の信頼性が評価できること、そして信頼性が低い地図であっても読み方を工夫すれば情報をきちんと取り出せることなどを示しました^[01]。

このように画像データは、文字情報を読むための材料となるだけでなく、非文字情報を「読む」ための材料ともなります。しかし非文字情報を読むという新しい読み方を実現するには、デジタル技術による支援、特に近年の発展が著しいコンピュータビジョンや機械学習の技術による支援が重要な役割を果たします。これらの技術は、問題を解くために必要な知識が少ないほど威力を発揮します。コンピュータビジョンは物理世界の性質さえ知っていれば解けるような問題に強く、機械学習は学習データさえ分析できれば解けるような問題に強いです。またこれらの技術は、いずれも数値の世界を対象とする点が共通しています。従って人文情報学が開発すべきデジタル研究技法は、数値の世界から有用な情報を取り出す部分を技術的に解決し、その成果を人間が活用しやすくすることで、人間が意味の世界をより深めていくことができる技法といえるでしょう。

本稿で紹介する「差読 (differential reading)」というアイデアもこの種の技法のひとつであり、文字史料を最初にあえて非文字的に「読む」ことによって、最終的に人間の読みの負担を軽減することを目指しています。

2. 差読とは？

まず中野による^{はんぼん}版本書誌学の解説書^[02/03]を参考に、江戸時代の出版について調べてみましょう。江戸時代の出版は、^{はんぎ}版木に文字や絵を彫って印刷する^{もくはん}木版印刷が主流でした。版(板)権という言葉が端的に示すように版木は財産となるほどの貴重なものであり、その作成には多大なコストを要したことから、修正が必要な場合でも「^{うめき}埋木」などによる部分的な修正にとどめるなど、新しい版木の作成をできるだけ避けていました。このような修正も含

め、版本の変異は、(1) 刊^{かん} (板・版)、(2) 印^{いん} (刷・摺)、(3) 修^{しゅう} (補・訂) の3つのタイプで区別します。刊とは新しい版本を彫って本を刊行すること、印とは既存の版本を使って本を刷ること、そして修は既存の版本に対して埋木などを使って部分的な修正を加えることを指します。

3種類の変異のうち、特に興味深いのは(3)のケースです。このタイプの変異を扱うには、ふたつの版を見比べて、その間に生じた部分的な修正を差分として検出する作業が必要となります。これは人間にとっては注意力を消耗するつらい仕事ですが、機械にとっては疲れ知らず(?)で連続的に実行できる得意な仕事です。このように、人間が苦手で機械が得意な仕事は、機械にやってもらうほうがよいのではないのでしょうか。このような「人機分業」を活用すると、(1) 機械が差分領域を検出する、(2) 人間が差分領域を読むという、新しい読み方を考えることができます。これが、われわれの提案する「差読」です。

3. 画像ベース差分検出

まず差読を実現するために必要な技術について考えてみましょう。テキストの場合、異なる版を比較するという研究には長年の歴史がありますが、まず画像からテキストを取り出さないとこの方法は使えません。それに対して画像をそのまま使う画像ベースの差分検出なら、テキストを取り出す必要はありません。2枚の画像をそのままマッチングし、重ね合わせ、画素単位の差分を検出することで、異なる版の比較が可能となります。

この中で、最初のステップとなる画像マッチングは、コンピュータビジョン研究における基本問題であることから、これまで非常に多くの研究が行われてきました。第一に、2枚の画像から特徴点と呼ばれる点を自動検出します。これは物体の境界や角^{かど}のように、物体の位置や角度が変わっても不変となる点が主に選ばれます。第二に、特徴点周辺の画素を分析し、特徴点に対応する特徴量ベクトルを算出します。このベクトルは、特徴点同士の類似性

を評価するために用います。第三に、2枚の画像が最もよく重なる画像変換パラメータを探索します。これは3次元空間における射影変換行列を推定する問題に帰着します。最後に、2枚の画像の画素値を比較し、画素値が大きく変化した画素を選び出すことで、差分検出は完了します。以上に述べた機能は、OpenCVなどのオープンソースソフトウェアでもすでに提供されており、スクリプト言語などから気軽に試してみることも可能です。

ただしこうした手法も万能ではありません。撮影条件が大きく異なる場合（例えばマイクロフィルムとデジタルカメラ）や、紙面が湾曲しているような場合は、マッチングの精度が大きく低下してしまいます。これから画像を撮影する場合は、できるだけ撮影条件を揃え、紙面をフラットにした状態で撮影するように注意してください。

4. 「武鑑全集」プロジェクト

最後に差読の適用例として、「武鑑全集」プロジェクトを紹介します^[04]。
藤實^{ふじざね}による解説書^[05]を参考にすると、「武鑑」とは以下のような特徴を備えた史料です。(1) 江戸時代に出版された大名家および幕府役人の名鑑^{めいかん}です。(2) 17世紀中ごろに出版され始め、慶応3年(1867)10月14日の大政奉還まで200年以上出版が続きました。(3) 実用書でありロングセラーブックでした。(4) 社会の需要に応じて、年を追うごとに厚くなり、その改訂の頻度は年に数度から月に数度までに増えました。われわれはこのように、時間方向に連続的に更新される多数のバージョンを有する史料を「時系列史料」と定義し、差読を活用した分析方法を確立したいと考えています。

現在の研究対象はROIS-DS人文学オープンデータ共同利用センター(CODH)が配布するオープンデータ「日本古典籍データセット」(国文学研究資料館所蔵)に含まれる381点の「武鑑」です(2017年12月現在)。特に須原^{すはら}屋茂兵衛^{やもへえ}が出版した『寛政武鑑』(寛政1年、1789)と『寛政武鑑』(寛政3年、1791)の2点を選び、画像ベース差分検出の有効性を確かめました。差分検

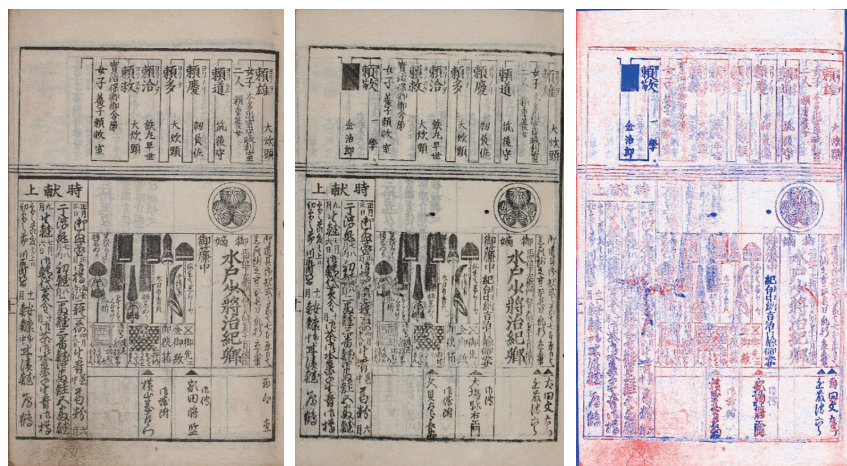


図1 「武鑑」の比較例。左が『寛政武鑑』(1789)、中が『寛政武鑑』(1791)、右が両者を比較した結果。1789年版のみ存在する部分は赤、1791年版のみ存在する部分は青で着色している(ただし白黒印刷の場合は判別できない)。

出ソフトウェアにはOpenCV 2.4 を利用し、特徴点検出にはFAST、特徴記述にはBRIEF、マッチングにはハミング距離(全探索)、射影変換行列の推定にはRANSACを用いました。そして重ね合わせ画像に対して、1789年版のほうが暗い画素は赤、1791年版のほうが暗い画素は青で着色することで差分をカラー強調し、差分が小さい画素は白で表示し背景化しました。その結果を図1に示します。この強調表示を見れば、1791年版では左上の系図に追加があること、右下の人物にも複数の追加や変更が存在することが一目瞭然です。

差読には以下のような利点があります。第一に、時系列史料に対して差分が生じた部分のみを^{ほんこく}翻刻(文字を打ち込んでテキスト化すること)する「差分翻刻」を適用することで、翻刻の手間が大きく減る可能性があります。第二に、埋木による修正だけでなく、非文字的な情報の変化(版木の部分的な欠損など)も検出できるため、テキストベース差分検出と比べて、バージョンの前後関係をより正確に同定できる可能性があります。第三に、「武鑑」における差

分は人びとの昇進や引退、死去などのイベントに対応するため、情報の変化率そのものが幕藩政治体制に関する新たな情報を提供する可能性があります。例えば災害は人事異動の頻度を高めたのか、といった新たな問いに答える道が開かれるかもしれません。

画像データを人間が読むためだけに使うのはもったいない。むしろそれを機械に読ませてみれば、文字が読めるという以上の新しい情報が得られるかもしれません。史料＝文字という固定観念から離れ、「非文字を読む」という発想を追追してみてはいかがでしょうか。

——注

- [01] 西村陽子・北本朝展「デジタル史料批判と歴史学における新発見」、『人工知能学会誌』Vol. 31, No. 6、2016年11月、769-774頁。
- [02] 中野三敏『書誌学談義：江戸の板本』岩波書店、2015年。
- [03] 中野三敏『和本のすすめ：江戸を読み解くために』岩波書店、2011年。
- [04] 北本朝展・堀井 洋・堀井美里・鈴木親彦・山本和明「時系列史料の人機分担構造化：古典籍『武鑑』を参照する江戸情報基盤の構築に向けて」、『人文科学とコンピュータシンポジウム論文集：じんもんこん 2017』2017年12月、273-280頁。<http://id.nii.ac.jp/1001/00184666/>
- [05] 藤實久美子『江戸の武家名鑑：武鑑と出版競争』吉川弘文館、2008年。