Vol.2012-CH-95 No.4
2012/8/4

近代文献のデジタルアーカイブ化とテキストマイニング —岩波書店「思想」を題材に

美馬秀樹[*]　丹治信[†]　増田勝也[†]　太田晋[*]

本研究の目的は，1921 年に創刊された岩波書店『思想』90 年分（約 1000 号，約 8600 論文，約 16 万ページ）を題材とし，電子化・構造化を行うことで，a)『思想』という知の集積，分析により 20 世紀日本の哲学・思想史を明らかにすること，b）分析結果の学部・大学院教育での活用の方法論構築を進めること，及び c）歴史的文献テキストの電子化，アーカイブ化に関する方法論を確立すること，である．本稿では，上記『思想』のデジタルアーカイブ化とテキストマイニングに関し，『思想』雑誌の電子化・構造化の手順とその問題点を報告し，特に，OCR による文字認識精度の向上，自動化・システム化に向けたレイアウト解析ソフトウェアの開発について，現状の取り組みと予備的に行った実験評価について報告する．

Digital Archiving and Text Mining of Modern-style Japanese Literatures using Iwanami Shoten's Journal "Shisou" (thoughts)

HIDEKI MIMA[†]　MAKOTO TANJI[††] KATSUYA MASUDA[††]　SUSUMU OTA[†]

The purpose of this study is to reveal Japanese modern history of philosophy by structuring Iwanami shoten's journal "Shisou" (Thoughts) using MIMA Search structuring knowledge system in which natural language processing (NLP), text mining and visualization are integrated.　Iwanami shoten's "Shisou" is one of the represent journals of philosophy in Japan, which has almost 90 years history from 1921 to the present-day, about 8,600 papers / more than 160,000 pages textual data. By digitalizing and analyzing the huge historical textual data using optical character recognition (OCR), NLP and the MIMA Search, we expect to discover new knowledge on Japanese historical flow of thinking during one of the most important eras, from before the World War II to the present-day.

1. はじめに

近年のデジタル化技術の発展と共に，文化的遺産のデジタル化保存に対する関心が高まっている．特に日本には，世界的に見ても稀なほど高度な人文学や芸術，文化創造の成果が集積さており，我が国の近代を通じた知と文化創造の蓄積は，21 世紀の人類の貴重な財産である．これら学術的，公共的知識資源をデジタル化し，蓄積，保存すると共に，アクセシビリティの向上等，高度に利活用可能な環境を迅速に構築する必要がある．

特にテキスト資源に関しては，ヨーロッパにおける大規模電子図書館プロジェクト（IMPACT）や，中国国家図書館での電子図書館プロジェクト（ほうせいプロジェクト），台湾，韓国等の漢字圏における同様の大規模電子図書館プロジェクトおいても，歴史的知識資産のデジタル化の取り組みが進みつつあり，アルファベットのみならず，漢字やひらがなといった 2 バイト文字文化圏での技術開発にも，大きな期待が寄せられている．日本においても，国立国会図書館により，このような公共的知識リソースを整備し，提供することが緊急の課題となっており，所蔵資料のデジタル化と全文テキスト化の実証実験が進められている．

東京大学知の構造化センターでは，上記のような文化的，公共的知識資源のデジタル化，高度な利活用技術の確立を目標とした，文理融合による文化的価値創出の研究を推進している．本稿では，そのパイロットプロジェクトとして進めている，岩波書店『思想』の構造化プロジェクトについて述べる．本プロジェクトでの目的は，1921 年に創刊された岩波書店『思想』90 年分（約 1000 号，約 8600 論文，約 16 万ページ）を対象とし，電子化・構造化を行うことで，a)『思想』という知の集積，分析により 20 世紀日本の哲学・思想史を明らかにすること，b）分析結果の学部・大学院教育での活用の方法論構築を進めること，及び c）歴史的文献テキストの電子化に関する方法論を確立すること，である．

以下，本稿では，上記『思想』の構造化プロジェクトに関し，『思想』雑誌の電子化・構造化の手順とその問題点を報告し，特に，OCR による文字認識精度の向上，自動化・システム化に向けたレイアウト解析ソフトウェアの開発について，現状の取り組みと予備的に行った実験評価について報告する．

2. テキストのデジタルアーカイブ化

通常のテキストデジタル化においては，あ）手入力，い）OCR（Optical Character Recognition）による文字自動認識，う）人手によるテキスト読み上げと音声認識によるテキスト化，が考えられるが，それぞれ現状では，テキスト

[*] 東京大学工学系研究科／知の構造化センター
School of Engineering / Center for Knowledge Structuring,, University of Tokyo
[†] 東京大学知の構造化センター
Center for Knowledge Structuring, University of Tokyo

化の精度，人的，金銭的なコスト，必要な時間に関し，Table 1 にあるようなトレードオフが存在する．尚，テーブルでの OCR とは Optical Character Recognition（光学的文字認識）を指している．これらを踏まえた上で，今後の完全自動化への技術開発が必要との判断より，本プロジェクトでは，スキャニング＋OCR によるデジタル化を選択することとした．

通常の OCR 処理では，

① スキャニング（イメージ化）

② OCR 処理による文字の認識

により，書籍や誌面からの文字情報の抽出を行う．現代の文献等に関しては既にいくつかのシステムが実用化され，99%以上の文字認識率が達成されている．しかしながら，膨大，かつ歴史的に貴重な資料を対象とした場合，デジタル化に係る時間や，古い文字（字体），言葉の認識等，様々な問題が生じる．また，スキャニングにおいては，フラットベッド・スキャナを利用したページ毎のデジタルイメージ化が一般的であるが，対象となる書籍の数が膨大であり，書籍に対し，人手によりページめくりを行いながらの作業は実用的ではない．さらには，歴史的に貴重な資料であるため，書誌背表紙の裁断による自動給紙システム等の利用にも難がある．そこで，本プロジェクトでは，米国キルタステクノロジーズ社（キルタス社）製のブックスキャナ装置を利用することとした(Fig. 1)．本ブックスキャナ装置では，ロボットアームによる自動ページめくりとデジタルカメラによる誌面スキャンが自動化されており，高速かつ，安全に書籍のデジタルイメージを作成することが可能である．

加えて，本プロジェクトの目標である，知の集積と高度な分析のためには，全文検索のみではなく，書誌データの抽出等によるリレーショナル・データベース化，及び構造化が必要である．そのため，

③ タイトル，著者等の書誌構造の認識，さらには，

④ 言語構造，意味構造の認識による，知の構造化

を行い，知の（再）活用を見据えたシステム化を行う．尚，④の詳細に関しては，本稿の範囲を超えるため，1),2)を参照されたし．

Fig. 1 キルタス社製 BookScanner（APT BookScan 2400RA）

Table 1 デジタル化手法の比較

	精度	コスト	時間
人手入力	◎	×	△
スキャニング＋OCR	○	○	◎
音声認識	△	△	×

(4 段階：◎，○，△，×)

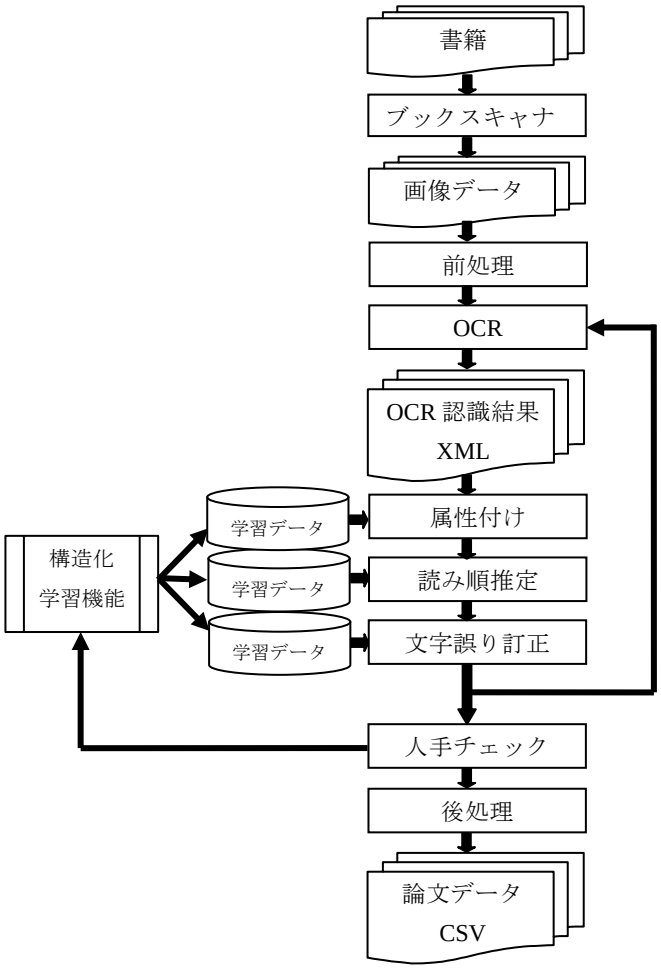

Fig. 2 書籍から論文データベース作成の流れ

総じて，本プロジェクトでは，書籍から構造化された論文データベースの作成に至る，以下のプロセスを設計し，実装を行った(Fig.2).

(1) ブックスキャナによるページ単位のデジタル写真撮影
上記ブックスキャナにより，書籍からのページ単位のデジタルイメージ化を行う．また，前処理として，画像処理による傾き補正，トリミング等の処理も同時に行う．

(2) OCR による自動文字認識，誌面レイアウト解析によるブロックの認識
OCR により文字認識を行う．その際に，文字領域，図形領域等の誌面上のブロックを自動で認識する．尚，本システムでは，OCR エンジンとしてメディアドライブ(株)製の WinReaderPRO を利用し，JIS 第二水準文字

までの認識を行うこととした．冒頭でも述べたが，現代文に対する文字認識に対しては，99%以上の精度による認識が可能であるが，旧字体の使用や，縦書き文書への横書きでの見出しやタイトルの挿入，縦書きアルファベットとの混在等，現在ではあまり用いられない特殊なフォーマットが存在したため，当初の認識率は 95%程度であり，実用には難があったが，精度向上への研究開発を進めることで，現状では約 98%の自動認識率を達成している．文字認識に関する実験，評価に関しては第 3 章で詳述する．

(3) 統計的文字誤り訂正
(2)で認識された文字に対し，統計的言語モデルおよび大量の OCR 結果から求めた文字類似度を用いて文字の誤認識の検出，および検出された文字誤りに対する自動文字誤り訂正を行う．文字誤り訂正に関する詳細は，第 3 章で詳述する．

(4) 誤認識文字の再学習
(3)で発見した誤認識文字，およびその文字に対する訂正文字の対を用いて OCR 再学習用のデータを作成し，(2)で利用した WinReaderPRO の学習機能へ投入することで，今後同様の文字が現れた際の認識精度を向上させる．

(5) 文字認識後処理 - 旧字体の新字体への変換
”藝術”等の現代語では通常使用しない文字を，旧字体→新字体への変換辞書により自動変換する．尚，厳密には旧字体と新字体では微妙な意味の違いが存在する可能性があるが，本プロジェクトでの議論の範囲を超えるとして扱わないものとする．

(6) 文書構造の認識－ブロックへの属性付け
通常，OCR により認識されるのは文字情報のみであり，また，書面におけるブロック単位での認識となるため，最終的に，論文等の文書の構造と共に結果を出力するためには，(2)で認識された文字の列とそのブロックへの自動属性付与が必要となる．本研究では，自動認識する属性として，タイトル，著者，ページ番号，注，本文等の意味的属性を設定し，さらに，論文区切り，発行年・月・号の取得も併せて自動認識することとした．これらに対する詳細と実験，評価に関しては同第 4 章で述べる．

(7) 文書構造の認識－読み順推定，
さらに本研究では，より精度の高い分析やアクセシビリティ向上に係る自動読み上げに対応するため，ブロック単位での読み順の推定を行う．(2)とも関連するが，昭和初期の紙が貴重であった時代背景等により，例えば，紙面節約のため，ページ途中で前の論文が終了した後に，改ページを行うことなく，次の論文が開始される等，読み順推定においても非常に複雑な構造解析が必要なケースがいくつか存在する（Fig. 3）．このため，段組が重なる場合等，通常の縦書き文書における右から左，上から下への読み順のルールのみでは，正確な推定ができない．本プロジェクトでは，まず，ルールベースの文書構造（ブロック属性と読み順）認識エンジンを開発した．これらに対する実験，評価に関しては第 5 章で詳述する．

尚，現状では，上記全てを 100%の精度で行うことは難しいため，適宜，人手による修正を行えるシステム構成とした．これらの自動認識と人手による修正を得て，最終的には，構造をもった書籍の情報を XML，もしくは CSV により出力することが可能である．

Fig. 3 複雑な構造認識，読み順推定が必要なレイアウト例

3. 文字認識と評価

各年代において論文一遍を任意に抽出し，OCR 認識結果に対して人手で誤り文字の箇所を数え上げ文字認識精度を求めた．各年代における文字認識精度は以下の通りである．

Table 1. 年代別文字認識精度

年	文字数	誤認識数	認識精度
2000	19558	31	99. 81%
1990	12496	30	99. 75%
1980	32229	83	99. 74%
1970	8232	46	99. 44%
1960	3090	61	98. 02%
1950	6006	170	97. 19%
1940	8316	52	99. 37%
1930	4294	348	91. 90%
1920	9218	1078	88. 31%

新しい年代においては，原本の状態もよく 99%台後半の認識精度となる．しかしながら古い年代に行くにつれ日焼けや汚れ等原本自体の状態，旧字体や現代とは異なるフォントの使用などにより，認識精度は 90%程度となってしまう．また，年代によらず，アルファベットなどの横倒し文字や，記号類は他の一般の文字に比べ総じて認識率が悪い傾向がある．

こうした文字誤りに対し，統計的言語モデルおよび大量

の OCR 結果から推定した文字類似度を用いて OCR における文字の誤りを自動的に検出・訂正する．文字誤り訂正は以下の 3 ステップにて行う．

Step1:文字誤り検出

既存研究 8)の手法と同様に統計的言語モデルを用いて OCR 結果中の特定の文字列が誤りであるかどうかを判定する．OCR 結果文字列において文字トライグラム確率値が閾値未満の箇所についてはそこが文字誤りであると判定する．

Step2:訂正文字候補生成

上記 Step1 で誤りであると判定された文字に対し，訂正候補となる文字集合を生成する．訂正候補文字としては，1)OCR システムが出力する出力結果以外の文字候補，2)統計的言語モデルを用いて生成した文字候補 の二種類を利用する．後者については，対象誤り箇所の周辺の文字と文字トライグラム確率を用い，誤り箇所に入る確率の高い文字候補の集合を生成する．

Step3:訂正文字選択

上記 Step2 で生成された候補から文字類似度と統計的言語モデルを用いて最終的に出力結果となる文字を選択する．具体的には，全候補文字列に対し，形態素解析システムの辞書を用いて単語列を生成し，その単語列に対する単語トライグラム確率及び候補文字と OCR 結果文字の類似度を組み合わせて最も尤もらしい候補文字列を出力結果とする．文字類似度は「同じ文字画像に対して OCR 文字候補として出力される文字は類似する」という考えのもと，大量の OCR に対する「同じ文字画像に対する OCR 文字候補としての共起確率」を用いる．

以上の 3 ステップからなる文字誤り訂正システム作成し，OCR 認識結果に対し誤り訂正実験を行った．実験では，古い年代の文献に対応するため，学習データとして青空文庫および「太陽コーパス」[9]を使用した．1940 年代の論文一遍に対し誤り訂正を行ったところ，正しい文字を誤った文字に変換してしまう誤訂正は起こるものの，全体としては精度の向上が見られた．

4. ブロックへの属性付けと評価

本項では，OCR 文字認識結果に対し，OCR 結果中の各ブロックに対して属性を付与する機構とその評価実験について述べる．付与した属性は，タイトル，著者，ページ番号，柱，本文である．

本研究では，文書構造の認識において，構造認識モデルとして，以下を特徴として用いることとした．

a) ブロック中の文字の（相対的）サイズ
b) ブロックの前後左右の余白
c) ブロックの（相対的）サイズ
d) ブロックのページ中の（相対的）位置

まず，これらの特徴に対し，それぞれの認識のための仮説として以下を設定し，それぞれを属性付与ルールとして実装，実験評価を行うこととした．

Ⅰ. 平均文字サイズの算出

全ページから平均文字サイズを求め，この値を font_size とする．

Ⅱ. ノイズ除去

font_size * NOISE_RATE 未満の行からなるブロックをノイズとみなし除去する．

Ⅲ. 上下のページ番号検出

上端と下端それぞれから最初に見つかったブロックの高さが font_size * PAGENUM_RATE 未満ならばページ番号とする．

Ⅳ. 左右の柱検出

ページ番号より下にあり，かつ，一番左(右)端のブロックを探す．見つかったブロック内の一番左(右)端の行を探す.font_size より大きい文字が 1 行中の 70% 未満の場合は柱とする．行の Y 座標がブロックの Y 座標よりも font_size * HASHIRA_RATE 以上から始まっているならば，その行を柱とする(字下げ対応)

Ⅴ. タイトル・著者候補検出

ブロックの平均文字サイズが font_size * TITLE_RATE 以上の大きさをタイトル・著者候補とする．さらにブロック内部の行単位で文字サイズをチェックし，1 行の全文字の 70%が以下の条件を満たす場合にタイトル・著者候補とする．

・漢字の場合は font_size * KANJI_RATE 以上
・カナの場合は font_size * HIRAGANA_RATE 以上

Ⅵ. タイトル・著者候補の精査

さらに，前ステップで得られたタイトル候補のうち以下の二通りの条件を満たすものをタイトル・著者とする．

・縦書きの場合は，ブロックの横サイズがページの横サイズの 1/3 より小さい
・横書きの場合は，ブロックの縦サイズがページの縦サイズの 1/6 より小さい

Ⅶ. タイトルおよび著者の決定

タイトル・著者のタグが付いているブロックのうち，座標がページの半分よりも下にあるものを著者として，それ以外をタイトルとする．

Ⅷ. サブタイトルを探す

タイトルと著者名に挟まれた行を探し，見つかった場合はサブタイトルとする．

Ⅸ. 上記全てに対し，条件を満たさないブロックを本文とする．

尚，上記ルールにおいて，実験では，経験的な値より，NOISE_RATE, PAGENUM_RATE, HASHIRA_RATE, KANJI_RATE, HIRAGANA_RATE に対し，それぞれ 0.35, 1.5, 3, 1.3, 1.5, 1.3 を用いた．

ただし，上記仮説による文書構造の認識は，現状では，

主に岩波書店『思想』の分析によるヒューリスティックな仮説から導出したルールベース解析が中心である．そのため，他の時代や分野の文書に対しても，同様の精度での認識ができない可能性もある．また，一般にルールベースでの解析には未知の事象に対する汎化が十分ではない可能性が高く，さらには，対応幅を広げるために，ルールを増加すればするほど，維持管理に係るコストが増す．時勢や分野を超えて，よりカバーレッジを広く，高い精度で認識を行うためには，ルール導出に関しても機械学習等を利用した自動化の仕組みを導入することが望ましい．よって，まずは先のルールベース属性付与と人手による修正の後，それら正解セットを学習データとして用いることで，機械学習ベースでの属性認識器を構築することとした．尚，機械学習ベースの手法においては，機械学習手法として Support Vector Machine(SVM)[3] を用い，以下の特徴量を利用することとした．

・ブロック位置
　ページ中での対象ブロックの位置座標(x 座標，y 座標)
・プロックサイズ
　ページに対する対象ブロックのサイズ(幅，高さ)
・平均文字サイズ(縦，横)
　対象ブロック中の全文字の平均サイズ
・余白長さ
　対象ブロックからみて,最も近いブロックまでの距離(上下左右，四方向)．ただし，最も近いブロックがない場合はページ端までの距離とする．(Fig.4 参照)

　また,より高度な認識を行うため,さらなる仮説として，言語的特徴がそれぞれの文書構造の認識に影響するものと仮定し，以下の特徴を追加利用することとした．これは，例えば著者名やタイトルには名詞の割合が多い等の特徴を仮定している．

・テキスト内「名詞」の割合
　形態素解析結果において品詞が名詞である単語の割合
・テキスト内「固有名詞（人名）」の割合
　同，品詞が固有名詞（人名）である単語の割合
・前ブロック平均文字サイズ
　対象ブロックより前で最も近いブロックの平均文字サイズ
・後ブロック平均文字サイズ
　対象ブロックより後で最も近いブロックの平均文字

Table 2. 文書構造認識の精度

属性	ルールベース手法		機械学習手法	
	適合率	再現率	適合率	再現率
タイトル	0.957	0.873	0.976	0.945
著者	0.987	0.945	0.996	0.974
ページ番号	0.997	0.992	0.994	0.996
柱	0.982	0.967	0.974	0.984
本文	0.979	0.992	0.993	0.992

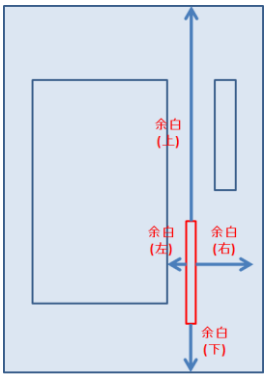

Fig. 4 余白長さの例

　最終的には，上記二種類の属性付与機構を構築し，岩波書店『思想』のデータに対しそれぞれの認識率について実験評価を行った．先にも述べたが，実験データはまず OCR 文字認識，行・ブロック認識を行ったあと，ルールベース手法で属性を付与した．その結果に対し，人手で修正を行うことで正解データを作成し，同じデータに対し機械学習手法で属性の付与を行い，両手法の結果と正解データを比較して適合率，再現率を求めた．なお，機械学習手法においては 10 分割交差検定を行い，両精度値を求めた．各値は以下のとおりである．

　上記のように，ルールベース手法においても一定の精度を確保することが可能ではあるが，これらと比較しても機械学習手法の方が全体的に精度が良いという結果となった．前述のように，機械学習では高い汎化による精度の確保が期待できる反面，全く異なるフォーマットに対しては，同様に学習用のデータを用意した方が精度の面では有利である．その点においても，ルールベース解析と機械学習の併用による一連の認識手法は有効であろう．

5. 読み順の推定と評価

　OCRで出力されたブロックの順序は，文章構造の観点から適切な順序ではない場合が多く，ブロック間の順序を決定する必要がある．また例えば論文誌では，本文，タイトル，著者情報などの論文の構成要素と，ページ番号や柱(ページ上部などにある章やタイトル情報) などの本のレイアウトに関する情報が混在するため，それらをそのまま繋げただけでは，音声読み上げなどでの問題がある．

　ブロックの読み順推定の既存件研究として，ブロックの論理ラベルと位置関係の論理演算によるルール[4]や，ページの分割による推定[5]，機械学習による手法[6]などが研究されている．しかし，学習による最適化が行えない場合や，日本語などの分かち書きのない言語に対応していないなどの問題があり，そのまま適用するのは困難である．

　ブロックは，それぞれページ中での X,Y 座標，及びブロックの属性（タイトル，本文などの論理ラベル）を持つ．既存研究5）のように,ページを分割していくことでブロックの順序を決定する方法を考える．

　Fig 5 左上のように一ページのブロックの上下方向，左

右方向に射影したヒストグラムを考える．ヒストグラム中の頻度が0 になる領域を分割線として上下，または左右に分割する（図では初めに縦方向に分割）．同様に分割された領域内で再度ヒストグラムを計算し，ブロックが一つになるまで再帰的に分割していく．この方法は，レイアウトの決まった書籍のような比較的シンプルなレイアウトで有効なヒューリスティックである．分割されたブロックからは階層的なブロックのツリー構造が得られる．非終端ノードにright-left, top-bottom のような予め書籍の種類により決められた選好ルールを適用していくことで，順序関係が得られる．（図の例ではA → B → C → D → E）．この分割方法は，単純なレイアウトでは有効に働くが，射影したヒストグラムが0 である部分がない場合や，逆に複数ある場合が考えられる．このような一般的な場合に対応するために，本研究では複数の分割候補を作り，その中でスコアの最も高い分割候補を次の分割として選択する方法を採用する．

スコアは次のように分割候補の特徴量ベクトル $\boldsymbol{X}=\{x_1, x_2, \ldots, x_N\}$の重み付き和で与える．

$$score(\mathbf{X}) = \mathbf{w} \cdot \mathbf{X}$$

ここで，特徴量としては以下のものを用いる．**W**は重みベクトルであり，この値は後述する方法で決定される．

- 分割が縦方向か，横方向か．
- 分割領域の射影されたヒストグラム頻度(図中の右，もしくは下のグレイ領域の高さ)
- 分割領域の中心線の位置
- 分割領域の幅
- ページモデルで分割領域にブロック境界が存在する頻度
- 分割しようとする領域にtitle, author ブロックが含まれているかどうか．

ここで，ページモデルとは，複数のページのブロック境界の頻度を取ったものであり，ブロックの切れやすさを表現するために，予め計算したものである．本研究では，ページを15 * 21 のグリッドに切り，その中にブロック境界が入る頻度を数えた．

分割候補は，縦と横方向のヒストグラムの各領域とし，このようにして算出されたスコアが最も高い分割領域で2つに分割する．同様に再帰的に分割を行うことで，一般の場合にも読み順の推定が可能である．本研究では重み係数ベクトル**w** の値を，汎用的な最適化手法である進化計算を用いて読み順推定ルールを学習する．特に，実数値ベクトルの最適化に用いられる差分進化(DE，Differential Evolution [7]) による最適化を用いた．

実験条件として，DE の繰り返し回数は30000 回，個体数は800 とし，データは「思想」の1940 年から1949 年の95 号分を用いた．また，「思想」データの80 号分をトレーニングデータとして用い，15 号分をテストデータとして用いた．正解データとの距離として，ブロックの順番のSpearman Footrule距離を用いる．

最適化の結果，初期状態で重みがランダムな状態では距離が0.6程度だったのに対し，最終的に正解データとの距離が0.04程度まで学習が進めることができた．この値は一号分（数十～100ページ程度）のページ内に間違って推定された読み順の数が数カ所程度であり，ほとんどのページで正しく読み順が推定されていることがわかる．

ここで述べた手法により，読み順ルールによる自動読み順推定を行うことができる．しかし，100 %の精度の順序推定を行うのは現時点では難しい．実際のデータの利用という観点からは，推定された順序関係を有効に使うことが可能である．

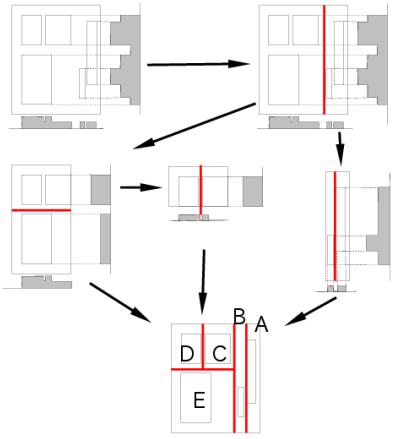

Fig. 5 ページの分割による読み順推定の例

筆者らが開発しているインタフェースをFig 6 に示す．左に大きくOCR 結果のブロックと推定されたページの順序が可視化されて提示される．ここでは，「思想」中の一ページを表示している．各ブロックはブロックの種類ごとに色付けされて表示されている．ブロック間の赤の矢印が推定された順序を表す．右のリストでは，ページ番号と段組の推定結果が表示され，ページ送りが行える．

また，本インタフェースでは，間違って推定された順序に対して人間がGUI 上で直接訂正を行うことが可能である．予備実験によると，人間が修正する時間は1 号(100 ページ程度) に対して5 分～15 分程度であり，平均は約7 分程度である．本プロジェクトでは約90 年分のデータをデジタル化することを想定しているため，この程度の時間は実用上問題ないものと考えられる．また，推定順序を提示しない場合は，平均15 分程度かかっており，推定順序の提示によって半分程度の時間で訂正を行うことができる．

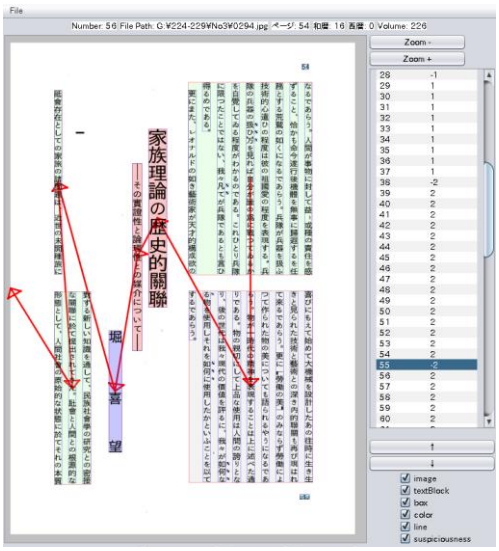

Fig. 6 訂正インタフェース

以上のデジタル化プロセス全体では，一号分（100～200 ページ程度）では概ね以下の時間で処理が可能である．尚，計測には最新の Core i7-3770K(Ivy Bridge)システム（メモリ 32GByte），OS として Windows7 を用いた．

1. スキャン画像の切り出しの前画像処理（約 10 分）
2. OCR によるテキスト認識と結果の XML 化（約 20 分）
3. レイアウト認識 （約 1 分）
4. 読み順推定 （約 1 分）
5. 文字誤り訂正 （約 30 分）

合計で約 60 分程度であり，90 年分のデータ全てを処理するのにかかる時間は，35 日～40 日程度と見積もれる．多くの処理は並列化によって同時に処理できることから，今後は並列化や処理の最適化による高速化を目指している．

6. MIMA サーチによるテキストマイニング

本研究では，上記により構築した『思想』90 年分のテキストに対し，分野や時勢を越えて関連する知識を抽出し，全体を俯瞰,再利用できる仕組みとして，テキストマイニング技術を利用したシステムを構築，利用することとした．以下が本研究によるテキストマイニング利用の目的である．

① 全体像の把握

知識の既存の関連や属性に基づく関連を抽出し，知識間の個々の関連から全体の関連を明らかにする．細分化や縦割りの弊害等により，失われがちな関連をも見つけ出すことが重要であり，オントロジー，可視化，見える化等の技術が重要な要素となる．

② 抽象化と詳細化

膨大な量の知識の全体像を把握するためには，抽象化は必須である．抽象化された領域より必要とする知識を選択した後，その領域の詳細化へと進めることで，必要な知識の絞り込みが容易になる．言わば，「森を見て，木を見る」操作である．

③ 合成

様々な知識から新たな知識を創造するためには，既存の知識を如何に再利用するかが重要である．異なる分野の知識を上記，抽象化等の操作により選択し，合成することで，より新しい知識の創出が期待される．また，創出された新たな知識を次の合成の種へとリサイクル，リファインメントを繰り返すことで，知識はより成熟する．

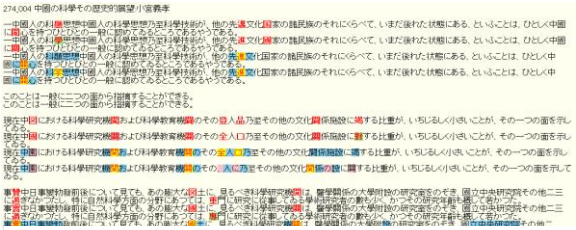

Fig. 7 重要表現の抽出

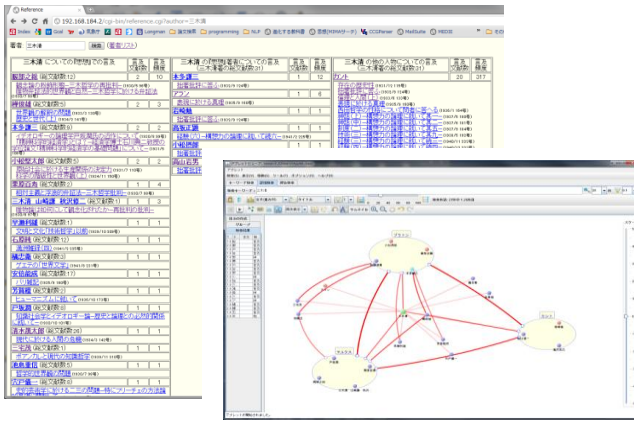

Fig. 8 著者間の関連の抽出と可視化

これらの分析，可視化，及び操作が，個々人，及び任意の視点によりリアルタイムに行えることが重要である．つまり，任意の視点で詳細化，抽象化の階層を上下しつつ，関連のある，もしくは関連が必要な知識を選択し，合成の要素を探すのである．さらには，次の瞬間にこれら新たに創出された知識が次の合成や抽象化の対象となる．このように，知識の連続的創出と活用を促し，さらに高度な知識の再活用へと昇華させるためには，知識創出，活用の「螺旋」を形成できることが重要である．

前述の目的を実現するためのシステムとして MIMA サーチ [1]を利用し，『思想』90 年分のテキストを実装したサイトを構築した．MIMA サーチは，用語抽出をはじめとした自然言語処理，テキストマイニング，及び可視化の技術を統合したシステムであり，既に東京大学授業カタログ [10]や，工学部シラバス，また特許等の検索，可視化システムとして実用化されている．MIMA サーチでは以下の機能を提供することが可能である．

- キーワードや年代等の属性指定による全文検索機能
- 検索された論文間の関連度の計算
- 上記により指定計算された関連を基に，関連の強い論文同士のまとめ上げ（クラスタリング機能）
- 上記のまとめ上げの任意の抽象度での実行（階層的クラスタリング機能）

MIMA サーチでは，これら分析結果に対し，グラフ構造による描画を利用した可視化が行えることが特徴である．

さらには，上記により検索された文書に対し，

- 重要な用語の抽出と可視化

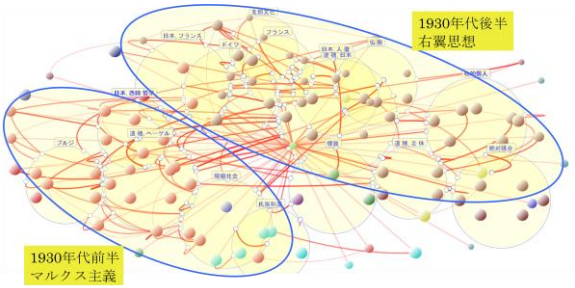

Fig. 9 “大和民族”に関する論文の分析と可視化

・構文解析を利用した頻出表現（係り受け関係）の抽出と可視化（Fig.7）
・著者間の参照の抽出と可視化（Fig.8）
・上記と他の属性（年代等）の指定によるクロス集計とその可視化を可能としている．

例えば，これらにより，指定の年代から論文を検索，可視化し，そこに含まれる用語や表現を抽出した上で，年代毎に集計，可視化するという一連の流れが簡単な操作で実現可能である．例えば，Fig.9 は，1930 年代の『思想』の“大和民族”に関する論文を分析，可視化したものであるが，トピックによる議論のクラスターが存在することが用意に見て取れる．また，Fig.10 は“研究”をキーワードとして，関連する論文に含まれる重要用語を年代別に集計し，グラフ化したものであるが，**「寺田先生」「物理学者」**が，1936 年の特集号を中心にヒット数を伸ばしているのに対して，**「自然科学」**は継続的に**「研究」**の重要語として論じられていたこと等を読み取ることができる．

7. まとめ

本稿では，東京大学 知の構造化センターで推進している文化的，公共的知識資源のデジタル化と高度な利活用技術の確立を目標とした取り組みの一つである，岩波書店『思想』の構造化プロジェクトにおける，デジタルアーカイブ化，及び MIMA サーチによるテキストマイニングについて述べた．

本プロジェクトでは，岩波書店『思想』90 年分の論文を対象とし，デジタル化，文書の構造認識，及び意味認識に至る，高度な分析，及びその可視化を行うことで，a) 20 世紀日本の哲学・思想史を明らかにすること，b）分析結果の学部・大学院教育での活用の方法論構築，及び c）歴史的文献テキストのデジタルアーカイブ化に関する方法論の確立，を目的とし研究開発を進めている．

メディアドライブ(株)製 OCR エンジンを利用した文字認識と，新たに開発した文書構造認識エンジンを統合することで，書籍のスキャンイメージから全自動でテキスト情報とその構造を出力するシステムの構築を行い，実際の『思想』90 年分を用いた実験と評価により，十分実用的に近代文献のデジタル化と構造化，及びテキストマイニングのシステム化が行えることを確認した．

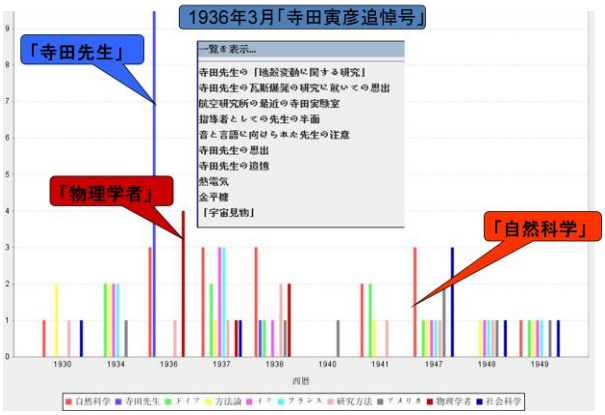

Fig. 10 重要用語の頻度の年代別集計とソート

今後は，デジタル化，及び構造化のさらなる精度向上を行う一方，これらを利用した学部・大学院での教育システムの構築を行う予定である．分野横断型教育の重要性が叫ばれる昨今において，文理の壁を取り払う新たな教育システム構築の一助となることを目指す．

謝辞 本研究に多大な協力を頂いた『思想』の構造化ワークショップ・メンバー，及び知の構造化センターRA の皆様に，謹んで感謝の意を表する．

参考文献

1) Mima. H., Ananiadou, S., Matsushima, K.: Terminology-Based Knowledge Mining for New Knowledge Discovery, ACM Transactions on Asian Language Information Processing, Vol. 5, (2006) pp.74-88.
2) 吉見俊哉，“コンピュータは思想史を書き換えられるか？ＭＩＭＡサーチによる 20 世紀日本の人文知への挑戦”，丸善ライブラリニュース，第 10 号, pp.4-5, 2010.
3) Corinna Cortes and Vladimir Vapnik. 1995. Supportvector networks. Machine Learning, 20:273–297.
4) Marco Aiello, Christof Monz, and Leon Todoran Combining linguistic and spatial information for document analysis. In Proceedings of RIAO ’ 2000 Content-Based Multimedia Information Access1, 2000.
5) Y. Ishitani. Document transformation system from papers to XML data based on pivot XML document method. Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings of ICDAR, 2003.
6) Donato Malerba and Michelangelo Ceci. Machine learning for reading order detection in document image understanding. Machine Learning in Document Analysis, pages 45-69, 2008.
7) Rainer Storn and Kenneth Price. Di_erential Evolution - A simple and e_cient adaptive scheme for global optimization over continuous spaces. Journal of Global Optimization, 11(4), 1997.
8) 竹内孔一,松本裕治，統計的言語モデルを用いた OCR 誤り訂正システムの構築，情報処理学会論文誌, Vol.40, No.6, p.2679-2689
9) 国立国語研究所，『太陽コーパス』(国語研究所資料集 15)，博文館新社
10) 東京大学授業カタログ, http://catalog.he.u-tokyo.ac.jp/ .