DEIM Forum 2021 G24-2

大規模収集した学術用語解説ウェブページ群の見易さ評定結果閲覧インタフェース

An Interface for Browsing Measurement of Visual Intelligibility of Large Scale Collection of Web Pages explaining Academic Concepts

曽田　耕生† 　大賀　悠平† 　岡田心太朗†† 　宇津呂武仁††† 　河田　容英††††

† 筑波大学大学院 理工情報生命学術院 システム情報工学研究群知能機能システム学位プログラム
〒 305-8573 茨城県つくば市天王台 1-1-1
†† 筑波大学 大学院システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1-1-1
††† 筑波大学 システム情報系 知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
理化学研究所 革新知能統合研究センター　〒 103-0027 東京都中央区日本橋 1-4-1
†††† (株) ログワークス　〒 151-0053 東京都渋谷区代々木 1-3-15 天翔代々木ビル 6F

あらまし　本論文では，ウェブ上で学術用語を解説するページ群を対象として，それらのページ群の見易さ自動評定結果の閲覧インターフェースの開発を行う．インターフェースでは，学術用語解説ページに対して，深層学習による見易さの自動評定を提示し，利用者によるページ選別の手助けを行う．また，特定の学術分野における学術用語，および，それらを解説するウェブサイト・ページを大規模に収集する手順を示し，閲覧インターフェースにおける学術用語の網羅性を高める．

キーワード　学術用語解説ウェブページ，閲覧インタフェース，見易さ評定，大規模用語収集，ResNet

1　はじめに

近年，学習の際に，多くの分野において学術用語解説ウェブページが存在し，それらを利用することによって，学術用語を学ぶ際の手助けとなる場合が多い．しかし，それらの学習コンテンツを探すために検索エンジンを用いる場合，学術用語を検索して上位に表示されるウェブページが，必ずしも学習に適しているとは限らない．そのような場合には，検索結果のウェブページの中から，学習に適したウェブページを探し出す必要があるが，この作業の労力を無視することはできない．そこで，文献 [3,4,8,9] では，学術用語解説ウェブページに対して，ウェブページの分かり易さや見易さに関する因子を分析するとともに，ウェブページのキャプチャ画像に対して見易さの自動評定を行う深層学習モデルを提案している．また，文献 [3,4,8,9] の成果を受けて，文献 [10] では，著者の一人が運営する「統計」分野のウェブサイトの各ページにおける，見易さ自動評定モデルの適用結果を分析しており，さらに，文献 [2] では，「統計」分野における用語解説ウェブサイト群を対象として，深層学習によって用語解説ウェブページ群の見易さを自動評定した結果をサイト横断的に閲覧するインタフェースを提案している．

以上の成果をふまえて，本論文では，特定の学術分野における学術用語，および，それらを解説するウェブサイト，用語解説ページを大規模に収集する手順を示し，閲覧インターフェースにおける学術用語の網羅性を高める．

2　学術用語解説ウェブページ群の大規模収集

本節では，「線形代数」分野を対象として，表 1 に示す規模の学術用語解説サイト，それらのサイト上の用語解説ページ，および，学術用語を大規模に収集する手順について述べる．

2.1　学術用語解説ウェブサイトの収集

「線形代数」分野を対象として学術用語解説サイトを収集する過程におけるサイト数の推移を表 1(a) に示す．

「線形代数」分野を対象として学術用語解説サイトを収集するにあたっては，まず，当該分野の代表的な学術用語 15 語 (クラメルの公式，クロネッカーのデルタ，ノルム，メネラウスの定理，ヤコビ行列，三角行列，二次形式，共役勾配，内積，対角化，正規直交基底，特性多項式，線形独立，行列式，写像) を対象として，各用語をクエリとするウェブ検索 [1] を行う．そして，各用語について，ウェブ検索結果の上位 30 ページに含まれるサイトを用語解説サイト候補として，計 220 サイトを収集した．220 サイトのうち，ウェブ検索結果の全 450 ページ中に 3 ページ以上のページが含まれるサイトに絞り込んだ結果においては，サイト数は 40 となった．40 サイトのうち，「線形代数」分野の用語解説ページが含まれないサイト，および，用語

1：Google 検索エンジン (https://www.google.com/) を用いる．

表 1　学術用語解説ウェブページ群の大規模収集: サイト数・用語数・ページ数 (対象分野: 「線形代数」)

(a) サイト数

	サイト群の収集段階	サイト数
(i)	15 用語 × 検索結果上位 30 ページ=450 ページから収集した候補サイト	220
(ii)	(i) のうち，全 450 ページ中に 3 ページ以上含まれるサイト	40
(iii)	(ii) のうち,「線形代数」分野の学術用語解説サイト	11

(b) ページ数

	ページの収集段階	ページ数
(i)	表 1(a)(iii) の 11 サイトの全ページ	14,776
(ii)	(i) のうち,「線形代数」分野以外の URL 情報を持つページを除外した残り	2,378
(iii)	(ii) のうち，表 1(c)(iii) の 1,575 用語の解説ページ (1 ページで 2 語以上を解説する場合を含むため延べページ数)	2,623
(iv)	(iii) のうち，1 文字の用語をページタイトル，または，サブタイトルに含むことが原因で，不適切もしくは重複するページとなったものを 2.2 節 (6) の手順により除外した残り (1 ページで 2 語以上を解説する場合を含むため延べページ数)	2,561

(c) 用語数

	学術用語の収集段階	用語数
(i)	表 1(b)(ii) の 2,378 ページから 2.2 節 (3) の手順で抽出した用語の候補	3,400
(ii)	(i) から学術用語以外を除外した残り	1,957
(iii)	(ii) のうちの「線形代数」分野の学術用語	1,575

解説の大半が PDF 文書に記載されているサイトを除外した結果，最終的に 11 サイトが得られた．

2.2　学術用語および解説ウェブページ群の収集

「線形代数」分野を対象として学術用語解説ページ，および，学術用語を収集する過程におけるページ数，および，用語数の推移を表 1(b)，および，表 1(c) に示す．これらの手順の詳細を以下に示す．

(1)　前節において収集した「線形代数」分野の学術用語解説サイト (11 サイト) 内の全てのページを収集した結果，14,776 ページが収集された (表 1(b)(i))．

(2)　(1) のうち，URL 中に `english`，`saiyo` 等の文字列が含まれ，他分野のページ，あるいは，当該サイトを運営する企業による採用情報のページ等,「線形代数」分野のページではないと判断できるページを除外した結果，2,378 ページが得られた (表 1(b)(ii))．

(3)　(2) のページにおいて，ページタイトル，および，サブタイトルを表す HTML タグである `title` タグ，および，`h1`∼`h6` タグのアンカーテキストとして用いられているテキストを収集し，そのテキスト中で片仮名・漢字のみで構成される文字列を学術用語候補として収集した結果，3,400 語の用語候補が得られた (表 1(c)(i))．

(4)　(3) のうち，学術用語でない用語 1,443 語，および,「線形代数」分野の用語でない用語 382 語を人手で削除した結果，1,575 語が得られた (表 1(c)(ii)(iii))．

(5)　(2) の 2,378 ページのうち，(4) の 1,575 語の解説ページの候補として，ページタイトル，あるいは，サブタイトルを表す HTML タグである `title` タグ，および，`h1`∼`h6` タグのアンカーテキスト中に (4) の 1,575 語を含むページを収集した．ここで，1 ページ中で 2 語以上の用語が解説される場合のページを重複して収集しているため，延べページ数は 2,623 ページとなった (表 1(b)(iii))．ただし，一つの用語解説サイト内に，同一の用語についての解説ページが複数掲載されている場合には，ページタイトルの文字数が最小のページを採用し，それ以外のページを除外する．

(6)　(5) のうち，1 文字の用語をページタイトル，または，サブタイトルに含む場合には，(a) そのページそのものが用語解説ページとしては不適切である場合，(b) 用語解説ページではあるものの，(4) で選定した 1,575 語の「線形代数」分野の用語からは外れる用語の解説ページである場合，および，(c) 1,575 語の「線形代数」分野の用語の解説ページではあるが，該当箇所そのものは，すでに別の用語の解説ページとして適切に選定済みである場合，が起こり得るため，そのようなページを除外する．ここでも，1 ページ中で 2 語以上の用語が解説される場合のページを重複して収集しているため，延べページ数は 2,561 ページとなった (表 1(b)(iv))．

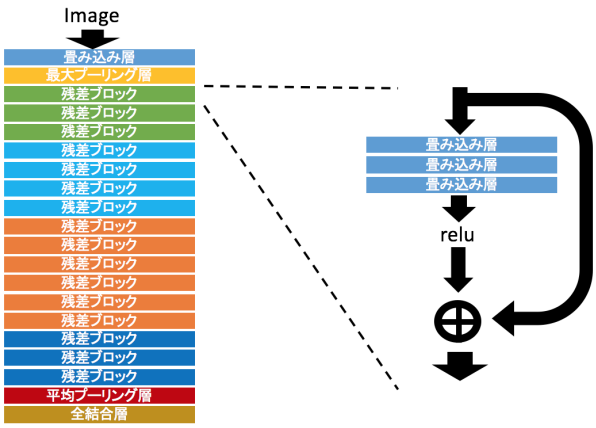


図 1　用語解説ウェブページの見易さ評定用 ResNet モデル

3　ResNet の fine-tuning を用いた学術用語ウェブページの見易さ評定モデル

3.1　ResNet

近年，画像認識の分野では，畳み込みニューラルネットワーク (CNN) を ImageNet などの大規模なデータセットに適用することにより，様々なタスクにおいて高い性能を達成している．また，ImageNet のような大規模なデータセットを用いて訓練した CNN のパラメータは，その汎用性の高さから，高性能な特徴抽出器として異なるドメインのタスクにおいても活用できることが知られている．そこで本論文では，ウェブページを画像化し CNN に入力することにより，ウェブページの見易さをふまえた全体評定を行う．その際，ResNet-50 モデル [1] を基盤の特徴抽出器として用いる．ResNet は，ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2015 [6] の「画像分類タスク (Image Classification)」,「物体検出タスク (Object detection)」，および,「位置特定タスク（Single-object localization）」において優勝したモデルである．このモデルは，49 層の畳み込み層と 1 層の最大プーリング層，平均プーリング層，および，1 層の全結合層 (1,000 値分類用) から成り立っている (図 1)．訓練済みの ResNet-50 としては，ImageNet2015 のデータセットによって 1,000 値分類用に訓練済みのモデルが一般に公開されており，この訓練済みモデルを fing-tuning することにより，他のタスクにも広く転用可能なことが知られている．本論文では，Python の深層学習ライブラリである Pytorch[2] 用に公開されているモデル [3] を利用して評価実験を行なった．学術用語解説ウェブページの画像に対する良否評定タスクにおいて ResNet-50 を適用する際には，ResNet-50 モデルの 1,000 値分類用の全結合層は使用せず，代わりに二値分類用の全結合層に取り替えたモデルを用いた．

2：https://pytorch.org/

3：https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py

3.2　訓練・評価手順

見易さ自動評定モデルの訓練・評価においては,「線形代数」・「解析」・「力学」・「電磁気」・「医学」・「IT 」・「生物」の各分野を訓練用分野,「化学」を開発用分野,「統計」を評価用分野とする．収集した学術用語解説ウェブページのトップ画面を画像化し作成したデータセットを用いて，ResNet-50 のパラメータを初期パラメターとして，fine-tuning を行う．20 エポック連続で開発データに対する正解率 (accuracy) が下降した場合に訓練を止め，開発データに対する正解率が最大となった訓練モデルを評価用モデルとして用いた．

4　Grad-CAM による見易さ評定理由の可視化

Grad-CAM (Gradient-weighted Class Activation Mapping) [7] は，CNN による評定理由を可視化するための手法の一つである．Grad-CAM の仕組みを図 2 に示す．Grad-CAM では，まず，畳み込み最終層の各特徴量マップに対して，自動評定モデルの自動評定値に対する寄与度 (勾配) を求める．そして，それらを重みとした特徴量マップの重み付き和を求め，この重み付き特徴量マップを元に画像化を行う．マップ要素値の正負によって「青」,「赤」それぞれのチャンネルに画素値を設定し，それぞれ正規化して画像化を行った．この手順により,「見易い」と判定された箇所は「青」で可視化,「見易くない」と判定された箇所は「赤」で可視化される．

5　大規模収集した学術用語解説ウェブページ群の閲覧インタフェース

「線形代数」分野における学術用語の解説ページ群を大規模収集したものを閲覧するインタフェースについて，その全体像および詳細を図 3 に示す．図 3 に示すように，この閲覧インタフェースにおいては，上部ヘッダー部分には各用語解説サイトへのリンク，左部サイドバーには学術用語が並んでおり，本体部分の各格子内には，それぞれの用語の解説ページへのリンクが提示されている．図 3 の学術用語解説ウェブページ群閲覧インタフェースを用いることにより，学習者が学びたい学術用語の解説ページの候補を探すことは可能となったが，例えば，初学者により相応しい「見易い」ページを厳選して提示することにまでは対応できていない．この点に関しては，文献 [2] で提案した「学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース」(次節，および，図 4～ 図 6 において詳述) を適用することによって，見易さ評定結果において高い評定となった用語解説ページを選択的に閲覧することが可能となる [4]．

4：次節においては，実装の都合上,「統計」分野を対象として「学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース」を動作させた事例を示している．ただし，本論文の設計においては，特定の分野を対象として，まず，本節で述べた「大規模収集した学術用語解説ウェブページ群の閲覧インタフェース」を動作させた後，そこから，同一分野の「学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース」を起動させることによって，二種類のインタフェース間をシームレスに行き来しながら，学習者が効率よく「見易い」用語解説ページにたどり着ける設計となっている．

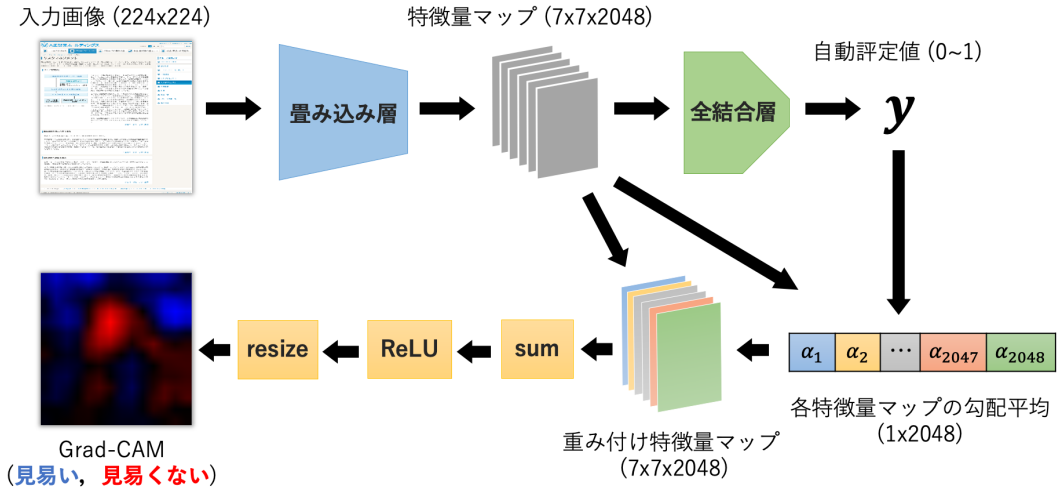


図 2　Grad-CAM による見易さ評定理由の可視化

6　学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース

本節では，文献 [2] で提案した「学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース」(図 4〜 図 6) について述べる．本節の例では，対象学術分野が「統計」分野の場合において，図 4 に示すように，当該分野においてよく知られた用語解説サイト名 [5] をインタフェース上部左右方向に配置し，用語が属する級 [6]，および，「統計」分野の学術用語を縦方向に配置する．そして，行列の各格子部分には，上部の各用語解説サイトから「統計」分野の学術用語 (「二項分布」,「回帰分析」,「仮説検定」,「観測値」,「ベルヌーイ試行」等)，をそれぞれ検索し，各用語のトップ画面の画像 (左半分)，および，自動評定理由を可視化した画像 (右半分) を左右連結したもの，および，見易さ自動評定モデルにより付与された見易さ確率の組を提示する．このとき，大抵の場合は，図 5 に示すように，各用語解説サイトの並び順は，各用語解説ウェブページに対する見易さ確率とは無関係な順となっている．したがって，このままでは，学習者が，見易さ確率が最大となる解説ページを掲載するサイトを探し出すために多大な労力を割くことになる．これに対して，本節の閲覧インタフェースでは，図 6 に示すように，同一の学術用語を解説する複数サイトのウェブページ群を見易さ確率の降順に整列する機能を設けており，この機能を用いることにより，学習者が，各サイトのページを見比べながら複数サイト間を自在に行き来することが可能となる．

7　関 連 研 究

本論文における学術用語解説ウェブページの見易さの自動評定に関する関連タスクとして，文献 [5] においては，本論文とほぼ同様の ResNet-50 モデルを用いた深層学習手法によって，プレゼンテーションスライドの画像情報に対する分かり易さを予測する手法を提案している．

8　お わ り に

本論文では，文献 [2–4, 8–10] の成果を受けて，特定の学術分野における学術用語，および，それらを解説するウェブサイト，用語解説ページを大規模に収集する手順を示し，収集した用語およびその解説ページをサイト横断的に閲覧可能なインタフェースを提案した．特に，本論文では，「線形代数」分野を対象として，実際に学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェースを実装することにより，多数の「線形代数」分野の用語とその解説ページを閲覧することが可能となった．今後の課題として，「線形代数」分野以外の多様な学術分野において本論文の大規模収集手順を適用するとともに，閲覧インタフェースを用いて，深層学習モデルによるウェブページの見易さ自動評定結果を提示することによって，閲覧インタフェースの操作性能の評価を行うことが挙げられる．

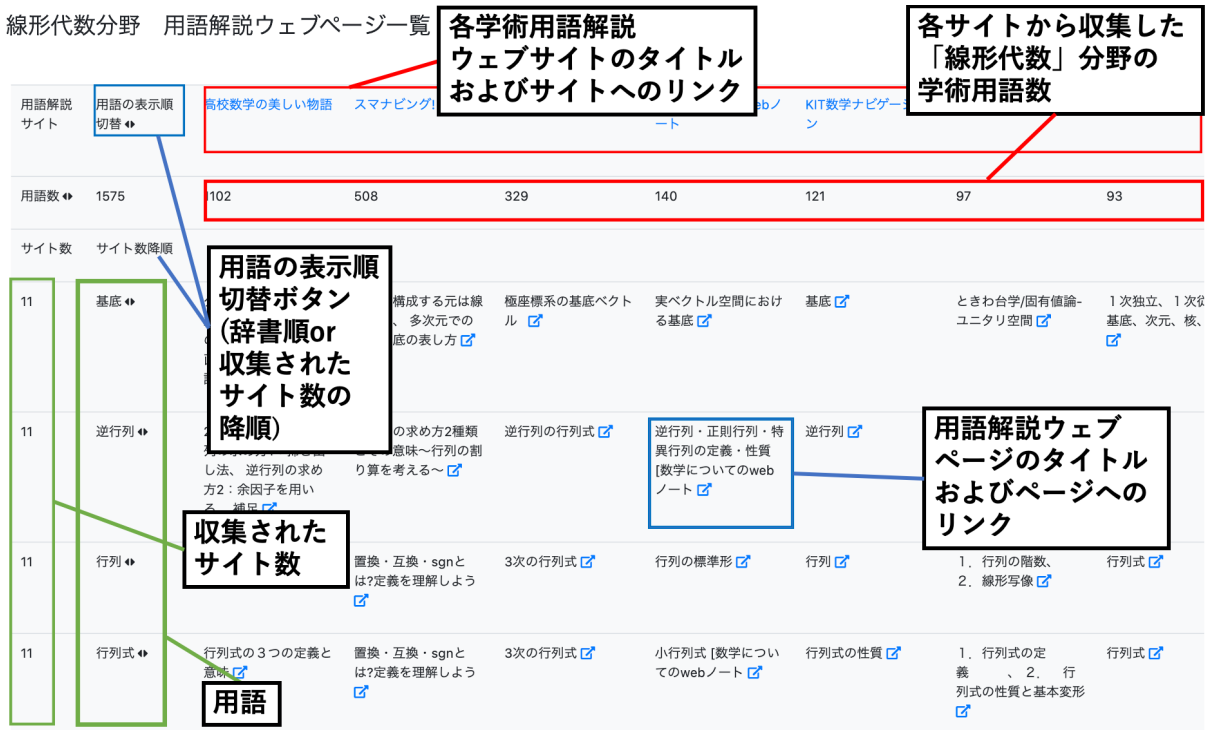
5：図 4 において列挙されている用語解説サイト名については，文献 [2] の著者の一人が運営する「旧: 全人類がわかる統計学」(「現: AVILEN AI Trend」(https://ai-trend.jp/) のみが実在したサイトで，他の 6 サイトについては，実在しない架空の用語解説サイト名を列挙してある．

6：「統計検定」(http://www.toukei-kentei.jp/) の各級を用いている．

文　　献

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pp. 770–778, 2016.
[2] 大川遥平, 曽田耕生, 岡田心太朗, 宇津呂武仁, 河田容英, 神門典子. 学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース. 第 34 回人工知能学会全国大会論文集, 2020.
[3] S. Okada, C. Hirohana, K. Kawaguchi, K. Soda, T. Utsuro, Y. Kawada, and N. Kando. Identifying factors of visual intelligibility of Web pages explaining academic concepts. In *Proc. DL4Ed*, 2019.
[4] 岡田心太朗, 塩川隼人, 韓炳材, 廣花智遥, 宇津呂武仁, 河田容英, 神門典子. 深層学習による学術用語解説ウェブページの見易さ自動評定結果の理由提示. 第 11 回 DEIM フォーラム論文集, 2019.
[5] 大山真司, 山崎俊彦, 相澤清晴. プレゼンテーションスライドの客観評価と印象予測. 第 16 回 FIT 講演論文集, 第 3 巻, pp. 45–52, 2017.
[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein,

(a) 全体像



(b) 詳細

図 3　大規模収集した学術用語解説ウェブページ群の閲覧インタフェース (「線形代数」分野)

A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252, 2015.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. ICCV*, pp. 618–626, 2017.

[8] 塩川隼人, 春日孝秀, 韓炳材, 宇津呂武仁, 河田容英. 深層学習を用いた学術用語解説ウェブページの見易さの自動評定. 第 10 回 DEIM フォーラム論文集, 2018.

[9] 塩川隼人, 岡田心太朗, 韓炳材, 廣花智遥, 宇津呂武仁, 河田容英, 神門典子. 深層学習を用いた学術用語解説ウェブページの分かり易さ・見易さの自動評定. 第 11 回 DEIM フォーラム論文集, 2019.

[10] 曽田耕生, 大川遥平, 岡田心太朗, , 廣花智遥, 宇津呂武仁, 河田容英, 神門典子. 学術用語解説ウェブページ見易さ評定モデルのサイト単位適用事例の分析. 第 12 回 DEIM フォーラム論文集, 2020.

図 4　学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース (「統計」分野) (1): 整列操作前の全体イメージ

図 5　学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース (「統計」分野) (2): 整列対象ページの整列前配置

図 6　学術用語解説ウェブサイト群の見易さ評定結果閲覧インタフェース (「統計」分野) (3): 整列対象ページの整列後配置