情報編纂システムのためのWebページ分割とその応用

A Web Page Segmentation Method for an Information Compilation System and its Applications

伊藤 太樹[*1] 柿元 宏晃[*1] 大囿 忠親[*1] 新谷 虎松[*1]
Taiki Ito Hiroaki Kakimoto Tadachika Ozono Toramatsu Shintani

[*1]名古屋工業大学工学研究科情報工学専攻
Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

A web page usually contains various blocks used for purposes such as navigation, decoration, interaction and contact information, which are not necessarily related to the topic of the web page. We focus on important blocks on Web pages to extract and reconstruct towards development of the Information Compilation System. We propose a web page segmentation method to extract important blocks. Our method consists of two points. The first point is to segment a web page in hierarchical fashion. The second point is to classify blocks into content types for sorting out. We have implemented applications applied to content blocks, and validated the utility of the proposed method.

1. はじめに

現在，Web は大規模な情報資源として重要視されており，Web ページを解析するための Web ページ分割研究が注目を集めている．本来，HTML におけるマークアップは，文書構造・レイアウトに対してのみ行われ，その内容，すなわち「意味的な」側面については，人間による読解が前提となっている．しかし，情報洪水時代と呼ばれる今日では，その全てを巨大なデータベースとみなす検索エンジンや，分散する資源を横断的に統合する情報サービスが不可欠なものとなっている．本研究では，情報をユーザにとって利便性の高い形に変換することを目的とし，それを情報編纂と定義する．情報編纂を実現するには，人間だけでなくコンピュータにも情報の「意味」を理解させる必要がある．

Web ページ分割は，Web ブラウザで閲覧した際に人が認識するコンテンツ（例えば，メニュー，ニュース記事，検索バー等）をシミュレートし，HTML を意味的な構造へと変換することが目的である．意味的なまとまりをブロックと呼ぶ．本研究では，新たな Web ページ分割手法として，階層的ブロック分割手法，およびブロックの役割分類手法を提案する．さらにブロックに基づく，情報の取捨選択や再構築，再利用を行う．本研究の応用として，携帯電話用記事閲覧システム，およびコンテンツ再配置システムを試作した．

以降，2 節に情報編纂システムと，そのための Web ページ分割について述べ，3 節で階層的ブロック分割手法，ブロックの分類手法について述べる．そして，4 節で応用システムについて述べ，本手法の有用性を示す．

2. 情報編纂のための Web ページ分割

本研究における情報編纂システムとは，Web ページから情報を取捨選択し，ユーザにとって利便性の高い形に変換するシステムと定義する．情報編纂システムとして，4 節で述べる携帯電話用記事閲覧システムやコンテンツ再配置システムが挙げられる．前者は，ニュースサイトの Web ページから記事のみを抽出し，閲覧性の高い形に変換するシステムであり，後者は，ユーザが自由に Web ページのコンテンツを再構成できるインターフェースを提供し，必要な情報のみを印刷可能なシステムである．

情報編纂を行うには，Web ページの各情報が有用であるか否かを判断しなければならない．そのためには，有用性が判断可能な単位に Web ページを分割する必要がある．一般的に Web ページは DOM 構造で表現でき，その情報の最小単位は葉ノードに対応する．しかし，葉ノード 1 つが持つ情報量は少なく，単体では有用性を判断できない．

本研究では Web ページ分割を用いて，Web ページに記載されている意味的なまとまりを 1 つのブロックとして抽出することを試みる．さらに，各ブロックの有用性を判断する基準には，ブロックが持つ意味を理解する必要がある．Web ページには，様々な意味を持つブロックが存在している．例えば，Web サイトのロゴなどを表示しているヘッダ，サイト内のコンテンツ一覧を表示しているサイトメニュー，記事の見出し，検索フォーム，広告などが挙げられる．しかし，意味的まとまりに関する情報は，HTML に陽に記述されておらず，レンダリング後の内容やレイアウトに依存するため，意味を直接理解することは困難である．そこで，ブロックを抽出するために，レンダリング後のレイアウト情報や DOM 構造を用いる．また，有用性を判断するために，ブロックの意味の代わりに，ブロックが Web ページ上で担う役割を用いる．役割とは，他の Web ページに遷移させる働きや，ユーザの入力を受ける働きのことである．

2.1 Web ページ分割手法

文献 [Hattori 2007, Baluja 2006] では，DOM 構造やレイアウト情報を用いた Web ページ分割手法が提案されている．文献 [Hattori 2007] では，各 DOM ノード間の DOM 構造上の距離に着目した分割ルールにより，DOM 木をブロックに分割している．文献 [Baluja 2006] では，各 DOM ノードの座標情報をパラメータとして，決定木を用いた機械学習によって Web ページを 9 つのブロックに分割する手法を提案している．

Web ページにおけるブロックの解釈は，人間ですら一貫性がないため，ブロックを一意に定めることは困難である．例えば，図 1 では，(a) 点線で示されるブロックと (b) 実線で示されるブロックの 2 通りが示されているが，どちらが正解とは言えない．また，(a) では印刷やブックマーク登録などのイン

連絡先: 名古屋工業大学大学院情報工学専攻，〒 466-8555 愛知県名古屋市昭和区御器所町，{taiki，kak，ozono，tora}@toralab.ics.nitech.ac.jp

図 1: ブロックの粒度

ターフェースも含んでおり，ブロックが大きいほど，複合的な役割を持つ事が分かる．つまり，ブロックには様々な粒度が考えられ，その粒度によってブロックが示す意味内容や役割が異なると言える．

本研究では，新たな Web ページ分割手法として階層的ブロック分割を提案する．階層的ブロック分割では，Web ページを粗く分割した状態から，徐々に細かくブロックに分割していき，それらを階層構造として保持する．ブロックを情報編纂に利用する際は，浅い階層のブロックを利用すれば，一度に多くの情報を取捨選択でき，逆に深い階層のブロックを利用すれば，細かく取捨選択できる．また，最も粒度が粗い最下層のブロックに対して役割分類を行うことで，役割を一意に定めやすくなる．

3. 階層的ブロック分割

3.1 ブロックの階層構造

本研究では，Web ページにおけるブロックの階層構造を図 2 のような 3 階層（1 階層：表示クラス，2 階層：コンテンツブロック，3 階層：最小ブロック）の木構造とする．表示クラスとは，Web ページを構成するヘッダ（V_h），フッタ（V_f），左メニュー（V_l），右メニュー（V_r），センター（V_c）の 5 種類のブロックであり，1 階層目では最大 5 種類のブロックに分割する．コンテンツブロックとは，表示クラスをより細かい意味内容に分割したブロックであり，サイトメニューやログインフォームなどがそれにあたる．最小ブロックとは，単体で役割を持ち得る最小のブロックまで分割した状態である．

4.1 節で述べる記事閲覧システムでは，主に 1 階層目，および 3 階層目のブロックを利用する．また，4.2 節で述べるコンテンツ再配置システムでは，主に 1 階層目と 2 階層目のブロックを利用する．このように，ブロックの階層を選択することにより，システムに適した粒度のブロックを利用することができる．

3.2 階層構造に基づくブロック分割

1 階層目では文献 [伊藤 2008] で提案した手法を用いる．これは，多くの Web ページがあるテンプレートに従って作成されているという仮定に基づいている．また，表示クラスが DOM 構造上の浅い階層で分離することが多いというヒューリスティクスを基に，DOM 構造，および各 DOM ノードのレンダリングされた位置や面積を利用して分割を行っている．

2 階層目では，HTML がレンダリング後のレイアウトに意味を持ち，隣接してレンダリングされた DOM ノードは，意味的にも類似しているという仮定に基づき，文献 [Ito 2008] で提案したブロック分割手法を用いる．この手法では，位置情報だけでなく背景色，枠，形，面積，DOM ノード名，DOM 構造上の関係などをパラメータとし，SVM を用いた学習によりコンテンツブロックを特定している．

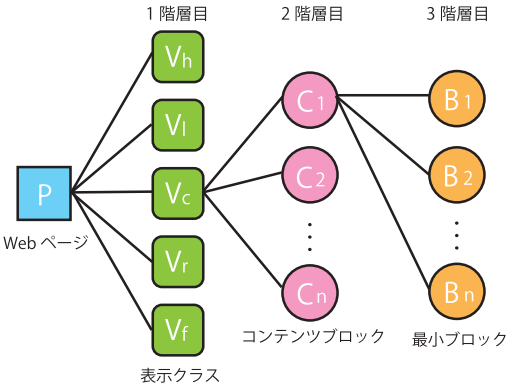

図 2: 階層的ブロック分割モデル

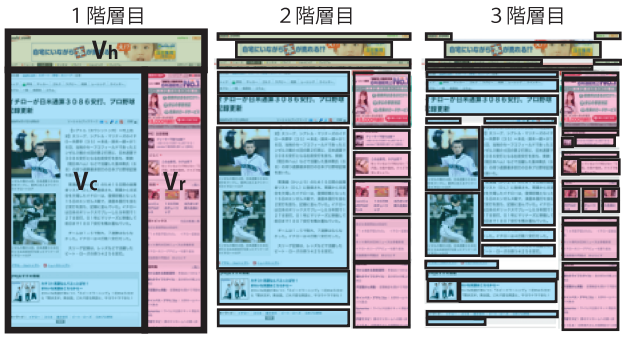

図 3: 階層的ブロック分割例

3 階層目では，Web ページの全体的なレイアウトが，入れ子構造になったブロック要素（DIV, TABLE, HR, P, H1 タグなど）によって決定されることに着目し，自身の子ノードにブロック要素を含まないブロック要素を最小ブロックする．HTML におけるブロック要素は，Web ページ上で矩形領域を確保し，内包する要素をその領域内に描画する．また，特殊なスタイルシートの影響を除けば，DOM 木解析を用いて互いに重なり合わないブロック要素の組を抽出することができる．

図 3 は Web ページを階層的に分割した例を示す．

3.3 ブロックの役割分類

Web ページには (1) 閲覧者に情報を伝える，(2) 他のページに導く，(3) 閲覧者と対話する，の 3 つの目的が存在する．例えば，ニュース記事や写真は (1) であり，サイトメニュー，関連記事一覧が (2)，印刷やブックマークなどの JavaScript を実行するボタンや検索フォームが (3) にあたる．本研究では，それらを役割の最小単位とし，それぞれ Information Class, Navigation Class, Interaction Class と定義する．

本手法では，初めに最小ブロックに対して役割分類を行う．3 つのクラスの中で，Interaction Class を抽出することは比較的容易である．最小ブロックがフォーム部品を持っている，もしくは href 属性や onclick 属性に javascript でアクションが記述してある場合，Interaction Class へ分類する．

Information Class と Navigation Class は，ハイパーリンクだけで他の Web ページに誘導するか，しないかの違いで分類される．例えば，Navigation Class は複数の整列したハイパーリンクによって他の Web ページへ誘導するが，Information Class は文章の中でハイパーリンクを用いて他の Web ページへ誘導する．そこで，文書とハイパーテキストの文字数比，最小

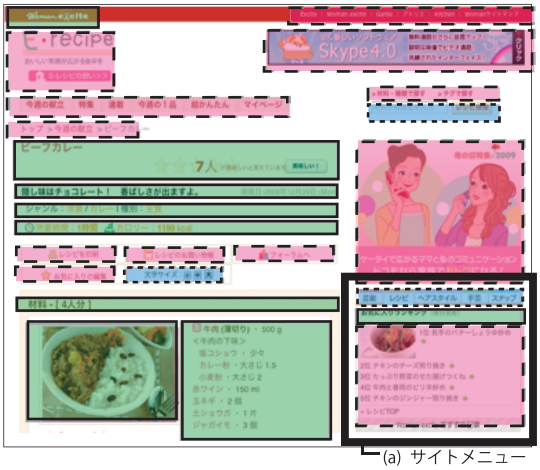

図 4: ブロックの役割分類結果

ブロックに対するハイパーリンクが占める面積割合，ハイパーリンク数などをパラメータとしたヒューリスティックスルールを用いて分類を行う．

ヒューリスティクスルール内の各パラメータは，機械学習を用いることで決定した．学習はニュースサイト，ショッピングサイト，レシピサイトなど計 30 ページ，3285 個の最小ブロックに対して手動で分類した結果を教師データとし，SVM を用いて行った．

その結果を用いて，実際に役割分類を行った例を図 4 に示す．図中の実線（緑）は Information Class，破線（赤）は Navigation Class，点線（青）は Interaction Class を示す．ただし，Flash によるコンテンツ表示はインターフェース，広告，メインコンテンツなど，様々な役割が考えられるが，本手法で特定することは困難であるため，Flash コンテンツは全て Navigation Class とした．

次に，2 階層目のブロックに対しても，役割分類を行う．2 階層目のブロックは，複合的な役割を持つが，内包する最小ブロックの役割と面積比を用いることで，そのブロックが持つ役割の割合が特定できる．例えば，図 4(a) サイトメニューのようなコンテンツブロックには，メニューを切り替える Interaction Class，メニューのタイトルを示す Information Class，リンクが複数個並ぶ Navigation Class に属する最小ブロックが存在する．しかし，それぞれの面積比で (a) の役割を Navigation Class と判定することが可能である．

4. 応用

4.1 携帯電話用記事閲覧システム

携帯電話に対する情報編纂システムとして，ニュースサイトの Web ページから記事だけを抽出し，操作性の高いインターフェースを持つ形式に変換するシステムを実装した．記事は，主に文章や画像から構成されており，本研究で提案した役割分類の結果が適用可能であることから，記事抽出への応用が有効であると考えられる．

本システムでは，Web ページに対して階層的ブロック分割，および役割分類を行い，3 つのクラスに分類されたブロックを得る．ここで，ニュース記事を構成する見出し，日付，写真，本文は全て Information Class に属する．また，ニュースサイトの多くが表示クラスのセンターに記事を配置しているというヒューリスティクスに基づき，センター以外の表示クラスを除去する．その後，センターに存在するコンテンツブロック

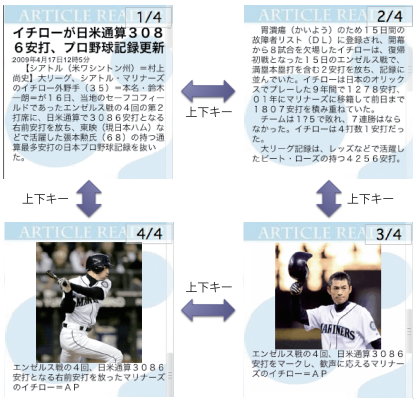

図 5: 携帯電話用記事閲覧システム例

図 6: コンテンツ再配置システム例

のうち，Information Class の割合が高いブロックを抽出する．最後に，抽出したコンテンツブロックが内包する Information Class 以外の最小ブロックを削除し，それを記事とする．

抽出された記事に対して，ページめくりなどの操作機能を付加して 1 つのファイルにコンパイルする．実際にニュースサイトで本システムを利用した例を図 5 に示す．図 5 では，1 つの記事が計 4 ページのコンテンツに変換されているが，上下キーを 1 回入力するだけで各ページ間を移動できるため，閲覧性の高いコンテンツに変換されていることが分かる．

4.2 コンテンツ再配置システム

Web ページには，ユーザが必要とする情報だけでなく，広告やロゴなど，Web ページ制作者側の意図により記載している情報も含まれている．そのため，Web ページを印刷する際に，ユーザにとって不必要な情報が含まれ，本来必要なコンテンツの表示領域が狭くなるという問題が生じる．そこで，PC 用の Web ブラウザに対する情報編纂システムとして，Web ページの各コンテンツをユーザが自由に再配置や削除できるインターフェースを実装した．

本システムでは，まず Web ページを階層的ブロック分割によってブロックに分割する．そして，図 6 のように，各ブロックごとに移動，削除，最小化，単体で表示などの再配置インターフェースを付加する．しかし，ブロックを移動させる，または削除する場合，レイアウトが崩れることがある．そこで，本システムでは，文献 [近藤 2009] のシステムを用いて，各ブロックを画像化する．

文献 [近藤 2009] のシステムは，Web ページを全て，もしくは部分的に画像化することで，どのような環境下でもレイアウトを崩す事無く表示することができる．また，JavaScript で仮想的にリンクを付与することで，通常の HTML 同様にリン

クを選択することが可能である．このシステムを利用することで，各ブロックを移動，削除した後もレイアウトを保ち，印刷時も表示されている状態をそのまま表現することができる．

各ブロックを再配置していく際に，最小ブロック単位で行うことはユーザにとって冗長な操作であり，表示クラス単位では細かい配置ができない．そこで，初期状態では表示クラスに再配置インターフェースを付加する．ユーザがさらに細かい再配置を行う際は，図 6(4) の細分化機能を選択する．細分化機能は，そのブロックを２階層目，３階層目と順に細かく分割したブロックに変換し，細分化されたブロックに再配置インターフェースを付加する．このように，操作するブロックの粒度を徐々に小さくすることにより，冗長な操作の削減できる．

ユーザが必要な情報のみを印刷する際，広告やナビゲーションリンク等が不必要となる場合が多い．そこで，Navigation Class や Interaction Class に属するブロックを一括で消す機能（図 6(5) より選択）も付加することで，単純な操作で情報の取捨選択が可能となる．

本システムの利用することにより，ユーザが必要だと考えるコンテンツのみを自由に再配置し，レイアウトを保ったまま印刷することが可能である．また，別の利用方法としては，(1) パーソナルポータルサイトの構築，(2) コンテンツの転載，(3) メール等に添付などが考えられる．

本システムでは，ブロックを画像化していること，JavaScript によりリンクが選択可能であることから，ブロック単体で表示してもレイアウトを崩さず，かつ本来の HTML としての機能を果たしている．そこで，利用頻度が高い Web サイト中のサイトメニューや検索フォームが含まれるブロックを１つの Web ページに並べることで，(1) のユーザにとって利便性の高いポータルサイトの実現が可能である．日記やレシピ，ニュース記事などを他のページに転載する際も，本システムのブロックを利用することにより，転載先のスタイルシートの影響を受けずに，本来のレイアウトを表現することができる．また，ブロックの画像だけを利用する事により，メールやデジタル文書などに挿入することも可能である．

このように，本システムを利用することで，情報の取捨選択だけでなく，容易に情報の再利用も実現できる．

4.3 評価

評価としてニュース記事領域の抽出精度を測定した．実験は，2009 年 2 月 1 日時点で Google ニュース[*1]に「冬」と検索した結果，上位 50 件の記事ページを対象とした．評価項目は，見出し，掲載された日時，本文，および写真がそれぞれ抽出できたかを，元の記事ページと比較する．見出し，掲載された日時，および写真は，抽出できたか，できなかったかの 2 項目で評価する．

表 1 に実験結果を示す．記事ページ数は，評価項目が記事ページ中に存在した数である．ニュース記事によっては，写真がないページ，日時が掲載されていないページも存在するため，この項目を用いた．評価項目の記事全体は，見出し，掲載された日時，写真，本文が全て抽出できた記事ページの数を表している．

実験により，4.1 節の記事抽出機能は，多くのニュースサイトの記事ページに対して有効であることが分かった．これは，全ての Web サイトにおいて，表示クラスのセンターに記事が含まれており，１階層目のブロックを利用することで，記事以外の情報の多くを簡単に削除できたためだと言える．ただし，掲載された日時のように，表示領域の小さなコンテンツは，最小ブロックとして抽出されず，Information Class 以外の役割に分類されることが多かった．そのため，抽出に失敗するケースが多く見られた．

∗1 http://news.google.co.jp/

表 1: ニュース記事抽出実験の結果

評価項目	抽出成功	記事ページ数
見出し	46	50
掲載された日時	36	44
写真	26	27
本文	42	50
記事全体	34	50

本手法の利点は，HTML 構造の異なりに影響されにくい点である．Web ラッパーを用いて記事抽出を行う場合，1 つの Web サイトに共通する目印を発見し，欲しい情報を抽出する．そのため，HTML 文書が少し変更されただけで対応できなくなる問題がある．本手法は，Web ページ分割し，その役割分類に基づいて記事領域を特定している．そのため，HTML 文書が変更された場合や，未知の Web サイトに対しても記事抽出が可能である

5. おわりに

本論文では，Web ページを意味的なまとまりである３階層のブロックに分割する手法，およびブロックの役割分類手法を提案した．また，その応用として，携帯電話用記事閲覧システム，コンテンツ再配置インターフェースを実装した．階層的にブロック分割することにより，ブロックの用途に合わせた粒度を利用することが可能である．さらに，ブロックの役割を特定することにより，使用するシステムに合わせて情報の取捨選択，再構築，再利用が可能となった．

参考文献

[Hattori 2007] Gen Hattori, Keiichiro Hoashi, Kazunori Matsui Matsumoto, and Fumiaki Sugaya : ”Robust Web Page Segmentation for Mobile Terminal Using Content-Distances and Page Layout Information.” , Proc. of the 16th international conference on World Wide Web, 2007.

[Baluja 2006] Shumeet Baluja : ”Browsing on Small Screens Recasting Web-Page Segmentation into an Efficient Machine Learning Framework.” , Proc. of the 15th international conference on World Wide Web, 2006.

[Ito 2008] Taiki Ito, Hiroyuki Sano, Tadachika Ozono and Toramatsu Shintani: “A Hierarchical Wew Page Segmentation Algorithm Using Machine Learing.” The Eleventh International Conference on Intelligent Systems and Control 2008, 2008.

[伊藤 2008] 伊藤太樹, 近藤圭佑, 浅見昌平, 大囿忠親, 新谷虎松: “携帯電話における Web コンテンツ閲覧のためのコンテンツ抽出アルゴリズムについて” 第 70 回情報処理学会全国大会, 2008.

[近藤 2009] 近藤圭佑, 浅見昌平, 大囿忠親, 新谷虎松: “マルチブラウザのための Web コンテンツの自動変換環境とその応用” 第 71 回情報処理学会全国大会, 2009.