ブロック解析に基づく Web ページの情報編纂システム

浅見 昌平 † 平田 紀史 † 大囿 忠親 † 新谷 虎松 †
名古屋工業大学大学院工学研究科情報工学専攻 †

1 はじめに

日々，World Wide Web では情報が更新され続けており，人々は情報過多によって必要な情報を見つけられない問題が起きている．特に，携帯電話からのインターネットアクセスが増加するに伴い，情報を集約する技術が求められている．情報洪水の中で，人々が必要な情報を見つけ出すための技術として，情報編纂が注目を集めている．情報編纂とは，情報を利用する人の必要性に応じて要約・組織化することである．情報編纂は，携帯電話のような資源が制約される環境において，人が必要な情報を得るための支援となる．

本研究では，既存の Web ページから広告などの不要な情報を除去し，情報を抽出するためのブロック解析手法を提案する．ブロック解析手法は，Web ページの特定位置に存在する情報を抽出する手法であり，情報編纂に利用可能である．情報抽出における既存の研究として Web ラッパーがあるが，ラッパー作成やメンテナンスのコストが高い．それに対し，ブロック解析手法はページごとに作り直す手間がかからず，レイアウトの位置指定による情報抽出ができる．本手法を用いた応用システムとして，携帯電話向けコンテンツを生成するための情報編纂システムを実装する．本システムでは，ブロック解析手法によってページ内の中央に存在する情報だけを抽出することで，Web ページの不要な情報を除去する．

2 ブロック解析

本節では，レイアウトや HTML 要素が持つ要素，属性を基に Web ページのブロック解析を行う．本稿におけるブロックとは，HTML におけるブロック要素が 1 つの矩形，もしくは 2 つ以上のブロック要素が結合された矩形を示す．

まず，Web ページ中の全てのブロック要素の中から，互いに重なり合わないブロック要素を検出する．互いに重なり合わないブロック要素とは，要素の中に他のブロック要素を含まず，絶対座標指定によってレイヤーが重なり合わないという条件を満たす要素である．このブロック要素を最小ブロックと呼ぶ．最小ブロックを検出することにより，Web ページを小さなブロックの集合に分割することができる．

次に，最小ブロックが持つ要素や属性の特徴から，3 つのクラスへ分類を行う．本研究で定義する 3 つのクラスを次に示す．

- Information Class: 文章や画像，Flash によって構成されるクラス
- Navigation Class: 整列されたハイパーリンクによって構成されるクラス
- Interaction Class: フォーム部品や JavaScript の呼び出し機能を持つボタンによって構成されるクラス

Information Class は，要素内に文章や画像，Flash によって構成されるクラスである．特徴として，それ自体が閲覧者に情報を提供する役割を持つことが挙げられる．例えば，写真付きのニュース記事や，ブログのエントリなどが該当する．

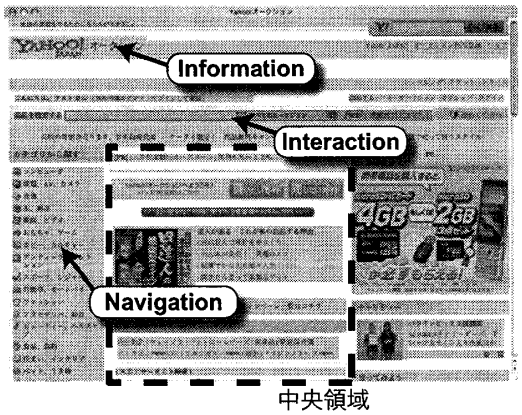

図 1: 最小ブロックの分類結果

Navigation Class は，整列されたハイパーリンクによって構成されるブロックである．整列されたハイパーリンクとは，Web ページに配置されている位置座標が縦方向，横方向，またはその両方向に並んでいるハイパーリンクを指す．このように整列されたハイパーリンクは，それ自体が情報を提供するものではなく，他のコンテンツへ誘導する役割を持っている．例えば，サイトのメニューバーや，ニュースヘッドラインなどが該当する．Interaction Class は，フォーム部品や JavaScript の呼び出し機能を持つボタンによって構成されるクラスである．その役割は，ブラウジングしている閲覧者が Web ページに対して行動を起こすためのインタフェースである．例えば，検索バーや，選択メニューが該当する．

一般的に，Web ページ内のブロックは大きくなるほど複合的な特性を持つ．例えば，最も大きなブロックは Web ページを内包する BODY 要素であるが，メニューや広告，記事など様々なコンテンツを持っている．しかし，本手法では，最小単位のブロックを検出し，その特性を最も良く示すクラスへの分類を試みている．1 つ 1 つのブロックが 2 つ以上のクラスに当てはまらないための工夫である．

最小ブロックを上記の 3 つのクラスへ分類した後，Web ページのテンプレート判別を用いて Web ページの中央領域に存在するブロックを取得する．テンプレート判別は，文献 [2] の手法を用いて Web ページをヘッダ，フッタ，右サイドバー，左サイドバー，および中央の領域へ分割する．ヘッダには，Web ページ上部に存在するサイトロゴやメニューバー，検索バーが含まれる．フッタにはサイトポリシー，右サイドバー，左サイドバーには縦方向のメニューバーや広告が含まれる．そして，中央領域には，Web ページの中央に存在する様々なコンテンツが含まれる．テンプレート判別によって，ブロックが配置されている Web ページ上の位置を判定できる．

図 1 に Yahoo!オークション[1]のトップページに本手法を適用して得られた最小ブロックの分類結果を示す．図では，ブロックのクラスに応じて色分けをしており，緑色のブロックが Information Class，赤色のブロックが Navigation Class，青色のブロックが Interaction Class へ分類されている．分類結果を見ると，ハイパーリンクが貼られていない見出しな

Implementing a Information Compilation System based on Block Analysis

Shohei ASAMI, Norifumi HIRATA, Tadachika OZONO, and Toramatsu SHINTANI
Graduate School of Engineering, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, 466-8555 JAPAN

[1] http://auctions.yahoo.co.jp/jp/

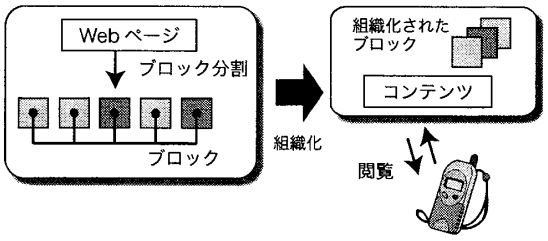

図 2: 情報編纂システムの概要

どが Information Class として分類されており，その内容は Navigation Class として分類される傾向にある．また，検索バーは全て Interaction Class として分類されている．また，点線で囲まれた部分は，テンプレート判別によって得られた中央領域である．

3 情報編纂システム

Web ページを携帯電話から閲覧する場合，画面の大きさ，操作機能の乏しさ，通信速度，または Web ブラウザの描画速度などの制約がある．本研究では，Web ページに含まれる有用な情報だけを組織化し，携帯電話から閲覧可能な情報編纂システムを構築する．情報編纂システムは，Web ページから抽出した有用な情報を掲載し，携帯電話での単純なキー操作によってページめくりが可能なコンテンツを生成する．

図 2 に情報編纂システムの概要を示す．まず，Web ページに対してブロック解析を行い，3 つのクラスに分類されたブロックを得る．次に，携帯向けに組織化するブロックを選択し，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．組織化とは，Web ページ内の不要なコンテンツを除去して，情報を探しやすくすることである．本研究では，Web ページテンプレートにおけるヘッダ，フッタ，およびサイドバーを除去する．

ブロック解析手法で得られたブロックの集合の中から，組織化するべきブロックを選択する．携帯電話向けコンテンツに組織化する場合，テンプレート判別によって得られた中央領域に存在するブロックを対象とする．コンテンツ上では，Interaction Class への操作は行うことができなくなる．そこで，組織化する候補の中から Interaction Class に該当するブロックを除外する．次に，メニューバーや広告などが含まれる Navigation Class ブロックを除外する．しかし，メニューバーや広告は文書中にラベル付けられていないため，経験的なルールを用いて除外する．広告は，外部ドメインのサイトへリンクされているブロックとし，メニューバーは 10 文字以内のテキスト，もしくは画像のみが整列されている Navigation Class とする．

図 3 に情報編纂システムによって組織化されたコンテンツを示す．Yahoo!オークションのトップページを本システムによって編纂した．ページの中央領域にある話題のキーワードのリンクや，その下にあるピックアップなどを閲覧することができる．通常，携帯電話に搭載されている Web ブラウザでページを閲覧すると，メニューバーやサイトロゴがトップにあり，目的の情報が探しにくい．本手法を用いると，ページの中央領域に存在する主要な情報を携帯電話で閲覧することができる．

4 評価

本研究で提案したブロック解析手法の評価を行う．ブロック解析では，Web ページ中のブロックの配置，およびその特徴を示すクラス分類ができる．そこで，12 の異なる Web ページに対して本手法を適用し，中央領域に存在するブロックのクラスを検証する．表 1 に実験結果を示す．ここでは，

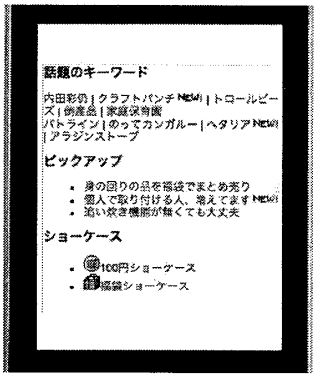

図 3: 携帯電話向けコンテンツの例

人が分類したクラスの結果と，ブロック解析で自動的に分類した結果が等しければ正しいとした．また，最小ブロックに分割するまでは，本手法を用いて分割している．

表 1: ニュース記事抽出実験の結果

クラス	分類数	正しく分類された割合
Information Class	356	87.6
Navigation Class	33	96.7
Interaction Class	14	85.7

実験では，Information Class に間違って分類されるブロックがいくつかあった．中でも，Navigation Class のブロックが誤分類されていた．Navigation Class に分類されたブロックは，広告やメニューバーが該当し，不要な情報の除去に役立つ．ブロック解析は，Web Wrapper のように抽出対象が明確なタスクと異なり，Web ページ上のレイアウト位置指定による情報抽出を行う．例えば，Web Wrapper では，日付情報を抽出するために「△月○日」という形式のデータを探し出すアプローチが考えられる．しかし，ブロック解析では，Web ページの特定の領域に存在する情報を抽出するというように，情報の記述形式が不明である．既存の Web ページ分割研究 [1] では，位置指定による抽出しか考えられていない．クラス分類によるブロックのラベル付けを行うことで，位置指定に加え，その特性による抽出対象の絞り込みができる点で，本研究は有用である．

5 まとめ

本研究では，Web ページのレイアウトや，HTML 要素が持つ要素，属性を基にしたブロック解析手法を提案した．ブロック解析手法では，ブロック要素が重なり合わないような最小ブロックを検出し，その最小ブロックが持つ特徴をもとにクラス分類を行った．また，テンプレート判別手法を用いてブロックが配置されている Web ページ上の位置を判定した．ブロック解析手法を用いた応用システムとして，携帯電話向けコンテンツを生成可能な情報編纂システムを実装した．本システムでは，Web ページの不要な情報を除去し，携帯電話から閲覧可能なコンテンツを自動生成できる．

参考文献

[1] S. Baluja: "Browsing on Small Screens: Recasting Web-Page Segmentation into an Efficient Machine Learning Framework," In the Proceedings of the 15th International World Wide Web Conference (WWW 2006), pp. 33-42, 2006.

[2] T. Ito, H. Sano, T. Ozono and T. Shintani: "A Hierarchical Web Page Segmentation Algorithm using Machine Learning," In the Proceedings of the Intelligent Systems and Control (ISC 2008), 2008.