階層的 Web ページ分割を用いたサブコンテンツ除去手法について

伊藤 太樹 †　柿元 宏晃 †　佐野 博之 †　平田 紀史 †
白松 俊 †　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

Web ページは，記事などのメインコンテンツが描画されている領域は一部であり，広告やメニューなどのサブコンテンツが大部分を占めている．閲覧者はメインコンテンツのみを閲覧，印刷，保存することが困難であり，Web Mining を行う際は，サブコンテンツがノイズとなり，精度低下に繋がっている．サブコンテンツを除去する一般的な手法として，サイト内の Web ページを収集し，出現頻度の高いコンテンツを除去する手法がある [1]．しかし，Web ページを収集するためのクローラー構築とクローリング時間などのコストが必要となる．本稿では，クローリングを必要としない手法として，コンテンツ構造に基づく除去手法を提案する．

コンテンツ構造とは，コンテンツという意味的なまとまりを 1 つのノードとした木型のネットワーク構造である．コンテンツが持つ意味を “内容的意味”，“配置的意味”，“機能的意味” の 3 つで定義する．内容的意味とは，コンテンツ内部にあるテキストや画像の内容から得られる意味であり，配置的意味とは，コンテンツが配置された場所によって生まれる意味である．例えば，インデントやマージンでコンテンツの階層や切れ目を示すことは，配置的意味である．機能的意味とは，HTML の機能によって生まれる意味であり，リンクやフォームのように特定の目的で配置されたコンテンツが持つ意味をさす．本研究では，配置的意味と機能的意味に着目したコンテンツ構造である “Block Net” を提案する．Block Net は，その構造とノード間関係からコンテンツの配置的意味を表現し，ノードが持つ特徴量から機能的意味を表現する．Block Net を用いることで，Web ページをコンテンツ単位で扱うことが可能となる．また 2 つの意味情報を利用することにより，コンテンツの取捨選択ができる．本稿では，Block Net におけるノードをブロックと呼ぶ．Block Net を構築するためには，Web ページをブロックに分割する必要がある．以下に，Block Net 構築のための Web ページ分割手法，及び Block Net を用いたサブコンテンツ除去手法について述べる．

2 Block Net の構築

2.1 Block Net

Block Net とは，ノード (ブロック b) を 1 つのコンテンツとしたネットワーク構造 $BN = (root, B)$ である (図 1)．図 1 の実線は，親子関係を表しており，各階層内の位置は，レンダリング時の位置関係を示す．$root$ はネットワーク構造の根を示すルートブロックであり，$B = \{b_1, b_2, \ldots\}$ は全てのブロックの集合を示す．通常の木構造とは異なり，任意の 2 ブロックは親，子の関係だけでなく，表示上の上下左右の隣接関係を保持する．例えば，図 1 の b_2 と b_3 は上隣接，下隣接の関係にある．任意のブロック $b_i \in B$ は，$b_i.doms$，$b_i.kind$，$b_i.param$ という 3 つの属性を持つ．$b_i.doms$ は，b_i を構成する DOM ノードの集合である．$b_i.param$ は，機能的意味に基づいた b_i の特徴量集合を示す．$b_i.kind$ は，b_i の種類を示す属性であり，属性値はテンプレートブロック，タイトルブロック，コンテンツブロックの 3 種類を定義する．

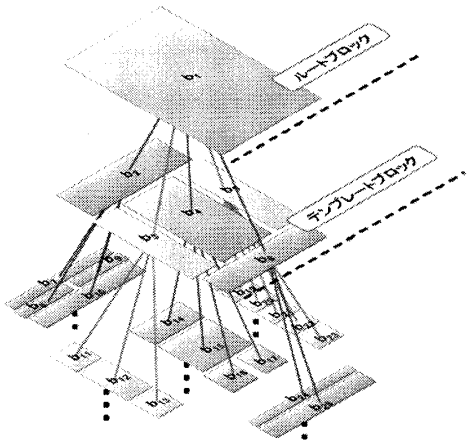

図 1: Block Net によるコンテンツ構造

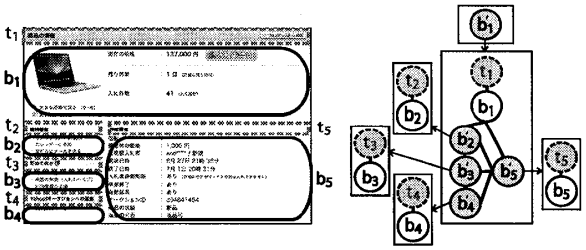

図 2: タイトルブロック

テンプレートブロックとは，Web ページの全体的なレイアウトを決定するブロックであり，本研究ではヘッダー T_h，フッター T_f，メイン T_m，サイドバー（左 T_l, 右 T_r）の 5 つを定義する．テンプレートブロックを他のブロックと区別することで，サブコンテンツの発見が容易になる．タイトルブロックとは，図 2 左の t_1 から t_5 のように，各コンテンツの見出しを示すブロックである．例えば，t_2 は b_2 の見出しであり，t_1 はその他全てブロックの見出しを表す．Block Net の構築には，Web ページをコンテンツ単位に分割することが重要になる．タイトルブロックを発見することにより，コンテンツ間の区切りやコンテンツの階層が明確になる．それ以外のブロックをコンテンツブロックと呼ぶ．

2.2 階層的 Web ページ分割

任意の Web ページ $page$ が与えられたときの，Block Net の構築手法について述べる．まず，$page$ をレンダリングし，得られた DOM 木と視覚情報から，$root.doms = \{< BODY >\}$ としたルートブロックを作成する．$root$ の子ノードには，文献 [2] の $ClassificationStep$ で抽出したテンプレートブロックを追加する．次に，各テンプレートブロックを DOM ノー

Eliminating Sub-Contents from a Web Page Using Hierarchical Web Page Segmentation

Taiki ITO, Hiroaki KAKIMOTO, Hiroyuki SANO, Norifumi HIRATA, Shun SHIRAMATSU, Tadachika OZONO, and Toramatsu SHINTANI

Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Instituteof Technology

ド単位のブロックに分割していく．分割する単位には，W3C (World Wide Web Consortium) が定義するブロックレベル要素 (DIV, TABLE, H1 など) を利用した．

最後に，タイトルブロックに基づき，複数のブロックを1つのコンテンツブロックに結合する．結合時は，タイトルブロックから下隣接方向に存在するブロックを子に持つブロック b' を追加する．図2左のWebページをコンテンツブロックに結合した例を図2右に示す．図中の四角は各階層であり，階層間の矢印の始点が親，終点が子を示す．また，各階層内の実線で結ばれた2ノードは隣接関係を持つ．図の例では，1階層目に t_1 を見出しとしたコンテンツブロック b'_1 を作成し，その子ノードに t_2 から t_5 を見出しとした b'_2 から b'_5 を配置する．このように，タイトルブロックを利用することで，階層的なコンテンツ構造を構築できる．

タイトルブロックであるかの判定は，決定木による分類器で行う．タイトルブロックは (1) 単純なHTML構造をしている，(2) 矩形領域は横に長く，下隣接ブロックと比較して面積は小さい，(3) 内部に含まれる文字数が少ない，の3つが挙げられる．この3つの特徴から，下位ノード数，縦横比，下隣接ブロックとの面積比，下隣接ブロックとの横幅差，文字数，HTMLタグ名等のパラメータを用いて，分類器を生成する．約300のWebページを対象に，人手で教師データを作成した際の交差検定では，95.4%と高い精度で判定可能であることが分かった．このことから，タイトルブロックであるかの判定は，十分に判断可能な問題だと言える．

2.3 機能的意味に基づく特徴量

Webコンテンツが持つ機能的意味には (a) 閲覧者に情報を伝える，(b) 他のページに導く，(c) 閲覧者と対話する，の3つが挙げられる．例えば，ニュース記事や写真は (a) であり，サイトメニュー，関連記事一覧が (b)，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 (c) にあたる．本研究では，それらをブロックの特徴量 $b.param$ に変換するために6個のパラメータを定義する．リンク以外のテキストノード面積と b の面積比 ($b.param_1$)，リンク以外の画像面積と b の面積比 ($b.param_2$) を (a) を表すパラメータとする．同様に，サイト内リンクの面積比 ($b.param_3$)，サイト外リンクの面積比 ($b.param_4$) を (b) のパラメータ，フォームの面積比 ($b.param_5$)，アクション付き要素の面積比 ($b.param_6$) を (c) のパラメータとする．

3 サブコンテンツ除去システムと性能評価

サブコンテンツ除去システムでは，入力されたWebページからBlock Netを構築し，その構造と特徴量を利用してサブコンテンツを除去した後のHTMLコードを出力する．ここで，メインコンテンツに対するヒューリスティクスとして (1)T_m にのみ存在する，(2) 情報を伝える役割 (2.3節の (a)) を持つ，(3) ページ上部に配置されている，(4) 1箇所にまとまって配置されている，の4つを挙げる．以上のヒューリスティクスから，ブロック $b \in B$ のメインコンテンツらしさを $f(b)$ で算出する．

$$f(b) = \begin{cases} 1 & (P(b) > \varepsilon) \\ -1 & (P(b) < -\varepsilon) \\ 0 & (otherwise) \end{cases} \quad (1)$$

$$P(b) = \alpha \cdot P_{func}(b) + \beta \cdot P_y(b) + \gamma \cdot P_{rel}(b) \quad (2)$$

$$P_{func}(b) = \sum_{i=1}^{2} b.param_i - \sum_{i=3}^{6} b.param_i \quad (3)$$

$$P_y(b) = \int_{b.y-T_m.y}^{b.y+b.h-T_m.y} g(y)dy \quad (4)$$

$$P_{rel}(b) = f(b.parent) + f(b.top) \quad (5)$$

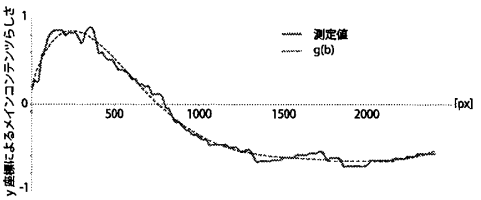

図3: y座標によるメインコンテンツらしさ $g(b)$

式中の $b.y$，$b.h$，$b.top$ は，それぞれ b のy座標，高さ，上隣接ブロックを示し，$T_m.y$ は T_m のy座標を示す．$g(y)$ は，b のy座標から算出するスコアで，事前に複数のWebページで測定した図3の測定値を近似式で表した関数である．P_{func} は，機能的意味を利用したスコアであり，P_{rel} は配置的意味を利用したスコアとなっている．Block Netを利用することで，これらの意味情報を容易に扱うことができる．

サブコンテンツの除去は，初めにテンプレートブロックのうち T_m 以外を除去する．次に，T_m の子ノードのうち，$f(b) = -1$ となるブロックを除去し，$f(b) = 1$ となるブロックは除去しない．$f(b) = 0$ となるブロックは，子ノードに展開して，再帰的に判定を行う．

サブコンテンツ除去システムの性能評価を除去精度＝(除去しなかったメインコンテンツの *HTML* コード量)/(メインコンテンツの *HTML* コード量)，除去率＝(除去したサブコンテンツの *HTML* コード量)/(サブコンテンツの *HTML* コード量) の2項目で測定した．対象としたWebページは，ニュースサイトとブログの2種類で，それぞれGoogleのニュース検索とブログ検索で「冬　雪」と検索した結果の上位20件ずつで行った．メインコンテンツを，ニュース記事もしくはブログ記事の見出し，記事，日付，写真，写真の説明とし，それ以外を除去すべきサブコンテンツとする．また，式1中の α，β，γ，ε をそれぞれ 1，1，0.5，0.5 とした．

評価実験の結果，除去精度は0.998，除去率は0.958となった．本システムをWeb Miningの前処理として，もしくは携帯電話などの閲覧支援として利用する場合，サブコンテンツを除去することよりも，メインコンテンツが残っていることが重要である．本実験では，99%以上の除去精度を保ったまま，約96%のサブコンテンツを除去できたことから，本システムの有用性は高いと言える．

4 おわりに

本稿では，Webページのコンテンツ構造であるBlock Netを提案し，その構築手法としてタイトルに基づく階層的なWebページ分割手法について述べた．Block Netは，ノードの構造，関係，特徴量で配置的意味と機能的意味を持つ．2つの意味情報を利用することにより，単純なヒューリスティクスルールのみで，96%程度のサブコンテンツの除去が実現できた．この結果から，Block Netが持つ意味情報は，有用性の高い情報だと言える．今後は，既存のWeb Mining手法から得られる内容的意味を，Block Netに加えることで，より人間が理解する意味に近い情報を抽出することを検討する．

参考文献

[1] Sandip Debnath, Prasenjit Mitra, Nirmal Pal, and C. LeeGiles: "Automatic Identification of Informative Sections of Web Pages.", IEEE Transactions on Knowledge and Data Engineering. 2005.

[2] 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.