3F1-4

確率モデルを用いたWebブロックの役割推定手法とその応用

Extracting Title-Blocks from Web Pages based on Probability Machine Learning

佐野 博之[*1] 白松 俊[*1] 大囿 忠親[*1] 新谷 虎松[*1]
Hiroyuki Sano Shun Shiramatsu Tadachika Ozono Toramatsu Shintani

[*1]名古屋工業大学大学院 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

Our Web page segmentation method divides a Web page into Smallest-Blocks, and then assemble some Smallest-Blocks into Content-Blocks. While smallest-Blocks have many roles, we focused on the title of Web contents. We adopted 9 parameters for each Smallest-Block in probability machine learning, and tried to obtain the extraction of Title-Blocks from Web pages. The experimental results show that Hidden Naive Bayes algorithm is suitable for extracting Title-Blocks from Web pages for our Web page segmentation method.

1. はじめに

本研究は Web ページ分割に関する研究である．Web ページ中に存在する，閲覧者にとって意味的にまとまりのある単位のことを，本研究では Web ブロックと呼ぶ．Web ページ分割とは，計算機によって Web ページを Web ブロック単位へと分割することである．本稿では確率モデルを用いて，Web ブロックからタイトルブロックを抽出する．

Web ページ分割アルゴリズムを確立することにより，様々な Web 技術の精度向上が期待できる．文献 [1] の中で Yi らは，Web ページ中からノイズとなるサブコンテンツ部分を除去した後に Web ページのクラスタリングを行うことにより，クラスタリングの精度が向上したことを報告している．Web ページの中にはメインコンテンツ以外にも，サイトロゴや広告，サイトメニュー，関連記事などのサブコンテンツが含まれることが多い．Web ページ検索システム，コンテンツフィルタリングシステム，情報抽出システム等でこのような Web ページを処理対象とする場合，メインコンテンツ以外のテキスト情報がノイズとなり，システムの精度が低下してしまう．システムの精度を向上させるためには，処理対象の Web ページを Web ブロックへと分割し，Web ページ中の主要な Web ブロックのみをシステムの処理対象とすればよい．

既存の Web ページ分割手法の多くが，Web ページを非常に細かい単位まで分割した後に，それらを一定のルールに基づいて意味的にまとまりのある単位まで結合している．本研究で用いる手法でも，Web ページを一度非常に細かいブロックまで分割する．結合時には，Web コンテンツのタイトルを表すブロック（タイトルブロック）に着目する．タイトルブロックに着目する理由として，ブロックに含まれるコンテンツ量に非依存な，汎用的な分割が行えるという点が挙げられる．

2. 関連研究

既存の研究で提案されている Web ページ分割手法には，大きく分けて，Web ページを記述している HTML の DOM 構造を用いた分割手法と，Web ページをレンダリングした結果のレイアウト情報を用いた分割手法の 2 つがある．

文献 [2],[3] では，DOM 構造を用いた Web ページ分割手法が提案されている．文献 [2] 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 ツリーをブロックに分割している．文献 [3] では，特定の DOM ノードを根とした部分木を生成し，そこから葉ノードまでのパスのエントロピーを用いて，Web ページ中の意味のあるブロックを抽出している．しかし HTML4 の特徴として，文章の内容と表現の分離が挙げられる．DOM 構造を持つ HTML ファイルには文章の内容のみを記述し，Web ページの見た目（表現）はスタイルシートに記述される．したがって，Web ページを閲覧者の観点から分割するためには，HTML の DOM 構造を解析するだけでなく，HTML をスタイルシートと共にレンダリングして得られるレイアウト情報も用いる必要がある．

文献 [4] では，VIPS アルゴリズムと呼ばれるレイアウト情報を利用したヒューリスティクスに基づく Web ページ分割手法が提案されている．フォント情報，面積，背景色，座標など，レイアウトに関する様々なパラメータを用いた 12 個のルールを HTML タグごとに使い分けることで，Web ページをコンテンツ単位に分割する．文献 [5] 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を行うことによって，Web ページを 9 つのブロックに分割する手法が提案されている．文献 [4][5] で提案されている手法では，Web ページを一度非常に細かいブロックまで分割した後，2 つのブロックにおけるフォントや背景色の違い，Web ページのレンダリング結果におけるブロックの面積やブロック間の距離などを利用し，フォントが同じである場合や距離が小さい場合に 2 つのブロックを結合している．しかし，それらフォントの違いやブロック間の距離が Web コンテンツの切れ目を明確に表している Web ページは少ない．また，長い文章の段落ごとには一定間隔の距離を空ける事も多いが，実際にはそれらの段落がまとまって 1 つの Web コンテンツを示している．ブロックの面積は，そのブロック内部に存在するテキストの量や画像の解像度などによって大きく変化する．そのため，同じ Web サイト内に存在する同一レイアウトの Web ページでさえも，メインコンテンツのテキスト量が変化した場合に，異なった Web ページ分割結果が作成されるという問題がある．本研究で用いる手法でも，Web ページを一度非常に細かいブロックまで分割するところまでは既存研究と同じであるが，その後にタイトルブロックに着目して結合を行う．タイトルブロックとは，下に

連絡先: 佐野博之，名古屋工業大学大学院 工学研究科 情報工学専攻，〒 466-8555 愛知県名古屋市昭和区御器所町，Tel:052-733-6550，Fax:052-735-5584，E-mail:hsano@toralab.ics.nitech.ac.jp

隣接するブロックの内容を説明するようなブロックのことである．タイトルブロックに着目した結合を行うことによって，ブロックに含まれるコンテンツ量に非依存な分割が可能になる．

3. Web ページ分割

3.1 Web ブロックの階層構造

本研究では Web ページは複数の Web ブロックから構成されるとする．Web ページ内で閲覧者にとって意味的にまとまりのある単位のことを本研究では Web ブロックと呼び，Web ページを Web ブロックへと分割することを Web ページ分割と呼ぶ．

Web ブロックには様々な粒度が考えられる．本稿ではテンプレートブロックとコンテンツブロックという 2 つの粒度に着目する．Web ページのレイアウトに着目して Web ページ分割を行うと，ヘッダーやフッター，サイドバーなど，非常に粗い粒度の Web ブロックへと分割できる．これらの Web ブロックは Web ページレイアウトのテンプレートを構成しているため，本研究ではテンプレートブロックと呼んでいる．テンプレートブロックは，ニュース記事やサイトメニューなど，より細かい Web ブロックから構成される．テンプレートブロックを構成する Web ブロックのことを，コンテンツブロックと呼ぶ．このように Web ブロックには階層構造が存在する．

本研究では，レイアウトを構成するヘッダー，フッター，レフトバー，ライトバー，センターの 5 つの矩形をテンプレートブロックとして扱う．Web ページをテンプレートブロックへと分割するためには，文献 [6] で提案した手法を用いる．我々は文献 [6] において，サポートベクターマシンによる学習を行うことによって Web ページのテンプレートを決定する手法を提案し，十分な精度が得られることを示した．

テンプレートブロックをコンテンツブロックへ分割するためには，ボトムアップ方式によって行う．すなわち，テンプレートブロックをいったん，細分化ブロックと呼ばれる非常に細かい単位まで分割した後に，それらの結合によってコンテンツブロックを生成する．

細分化ブロックへの分割には，W3C が定義するブロックレベル要素を用いる．ブロックレベル要素は Web コンテンツの配置やまとまったレイアウトを指定するために使われることが多い．ブロックレベル要素は Web ページ上で矩形領域を確保し，子供の要素をその領域内に描画する．Web ページの全体的なレイアウトは，入れ子構造になったブロックレベル要素によって決定される．本稿で提案する手法では，子ノードとしてブロックレベル要素を持たないブロックレベル要素を 1 つの細分化ブロックとして抽出する．ただしインライン要素であっても，細分化ブロックの兄弟ノードである場合には，そのインライン要素も 1 つの細分化ブロックとして抽出する．これにより，Web ページ上にレンダリングされる全ての要素がいずれかの細分化ブロックに属することとなる．

最後に，細分化ブロックを結合してコンテンツブロックを生成する．細分化ブロックの中でも特に，直下の Web コンテンツのタイトルを表すようなブロック（以下，タイトルブロックと呼ぶ）に着目した結合を行う [7]．Web コンテンツが多数配置されている Web ページには，人が閲覧したときに読解しやすいように Web コンテンツの上部にタイトルブロックが配置されていることが多い．すなわち，タイトルブロックは複数の Web コンテンツ間の仕切りとして利用することが可能であると言える．

表 1: タイトルブロック判定に用いる特徴量

特徴量	詳細
A_1	テキストノード長
A_2	テキストノードの面積 / ノード全体の面積
A_3	画像ノードの面積 / ノード全体の面積
A_4	ブロックの横幅 / 高さ
A_5	下隣接ブロックの面積がノードの面積より大きいかどうか
A_6	H1, H2, H3, H4, H5, H6，DT タグかどうか
A_7	同じ HTML タグが上隣接方向に連続している数
A_8	同じ HTML タグが下隣接方向に連続している数
A_9	下位 DOM ノードの合計数

3.2 タイトルブロックの判定

提案手法ではタイトルブロックに着目した細分化ブロックの結合を行うことによって，Web ページを Web コンテンツ単位へと分割する．そのためには細分化ブロックの中からタイトルブロックを抽出する必要がある．機械学習によってタイトルブロックを抽出するための分類器を生成する．タイトルブロックのレイアウトに基づく特徴として，以下の 4 つに着目した．

1. 下位 DOM ノード数が少ない
2. 内包する文字数が少ない
3. 高さに比べて幅が広い
4. 下に配置されているブロックより面積が小さい

上記の特徴の他に，HTML タグ名に基づく特徴量と DOM 構造に基づく特徴量を導入し，細分化ブロックがタイトルブロックに属するか否かの判定を行う．表 1 に示す 9 つの特徴量を用いた．以下に，特徴量の簡単な説明と，これらの特徴量を導入した理由を述べる．

A_1 から A_5 はレイアウトに基づく特徴量である．タイトルブロックは下隣接ブロックの内容を簡潔に表すキーワード・文章で構成されるため，テキストノード長は短くなり (A_1)，ブロック内部でテキストノードの占める面積の割合が大きくなる (A_2)．同時に，画像が占める面積の割合は小さくなる (A_3)．画像をほとんど含まず主にテキストノードで構成されるため，タイトルブロックは高さに比べて横幅が大きくなる (A_4)．タイトルブロックは下隣接ブロックの内容を簡潔に表すキーワード・文章であるため，下隣接ブロックよりも面積が小さくなる (A_5)．A_6 から A_8 は HTML タグ名に基づく特徴量である．H1,H2,H3,H4,H5,H6 タグは見出しを記述するために定義されたタグである．また，DT は Definition Term の略であり，DD タグとセットで利用される．DT タグの中に定義語を記述し，DD タグの中にはその用語の説明を記述する．つまり，DT タグは DD タグに記述した内容のタイトルを表していると言える (A_6)．上下隣接方向に同じ HTML タグが連続することは，そのブロック自身が隣接するブロックと並列関係にあることを意味する．タイトルブロックは直下に存在するコンテンツの内容を表すブロックであり，タイトルブロック自身が連続して出現することはない．したがって，同じ HTML タグが隣接して連続する可能性は低い (A_7, A_8)．A_9 は DOM 構造に基づく特徴量である．タイトルブロックは背景色やフォントで装飾するだけの HTML で記述される傾向にあるため，タイトルブロックが持つ DOM ノードの下位ノード数は少なくなる．

これらの特徴量を用いて機械学習を行い，タイトルブロックの分類器を作成する．本稿では確率モデルに基づき，Naive Bayes(NB)，Hidden Naive Bayes(HNB)[8]，Weightily Averaged One-Dependence Estimators(WAODE)[9] の 3 つの

アルゴリズムによって分類器を作成した．NB アルゴリズムを用いた理由は，非常に単純なモデルであるにも関わらず，訓練データが少なくてもうまく動作することが知られているからである．NB アルゴリズムでは各特徴量の条件付独立が仮定されている．しかしタイトルブロック判定のために用いる特徴量は独立ではない．例えば特徴量 A_1 はテキストノード長に関する特徴量であり，特徴量 A_2 はノード中でテキストノードが占める面積に関する特徴量である．これらの特徴量には依存関係が存在する．特徴量変数群の独立関係を緩和するアルゴリズムを用いるのが好ましい．特徴量変数群の独立関係を緩和する手法としては，HNB アルゴリズムや Averaged One-DependenceEstimators(AODE) アルゴリズムが知られている．HNB アルゴリズムでは各特徴に対して隠れた親を持たせることにより，特徴間の依存関係を表現することが可能となっている．AODE アルゴリズムは NB の変数間に対してリンクを追加し，複数の分類器をモデル平均するものである．Weightily Averaged One-Dependence Estimators(WAODE) アルゴリズムでは，単純なモデル平均ではなく，適切な重み付けを行い平均する．

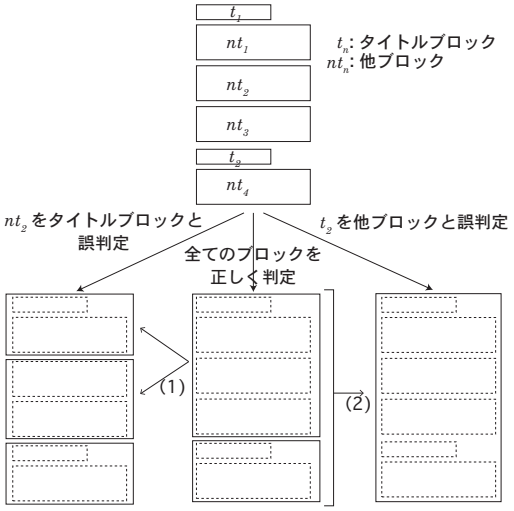

図 1: 誤判定によって意図しないコンテンツブロックが生成される例

4. 実験・考察

4.1 実験内容

本論文で提案した Web ページ分割手法ではタイトルブロックを用いた細分化ブロックの結合を行っており，タイトルブロックの判定精度が分割手法の精度を左右する．タイトルブロック判定精度を分類器生成時の 10 分割交差検定で測定する．

実験対象とした Web ページは，Google で「夏」と検索した結果の上位 50 件である．実験対象の Web ページから抽出された細分化ブロックは，全部で 8435 個あった．人手によってそれらのブロックをタイトルブロックとそれ以外のブロック（他ブロック）へと分類したところ，タイトルブロックが 836 個，他ブロックが 7599 個であった．他ブロックからランダムにサンプリングを行い，タイトルブロック 836 個，他ブロック 874 個の合計 1710 個のブロックを実験データとして用いた．

評価基準は，タイトルブロックを正しく判定した数 (a)，タイトルブロック以外のブロックを正しく判定した数 (b)，タイトルブロックを他ブロックと判定した数 (c)，他ブロックをタイトルブロックと判定した数 (d) の 4 つで行う．また，タイトルブロックの判定精度 P_{tb}，他ブロックの判定精度 P_{ntb}，タイトルブロックの再現率 R_{tb}，他ブロックの再現率 R_{ntb} を以下の式で求める．

$$P_{tb} = \frac{a}{a+d} \quad (1)$$

$$P_{ntb} = \frac{b}{b+c} \quad (2)$$

$$R_{tb} = \frac{a}{a+c} \quad (3)$$

$$R_{ntb} = \frac{b}{b+d} \quad (4)$$

また，それぞれの F 尺度 F_{tb}，F_{ntb} を以下の式で求める．

$$F_{tb} = \frac{2 \cdot P_{tb} \cdot R_{tb}}{P_{tb} + R_{tb}} \quad (5)$$

$$F_{ntb} = \frac{2 \cdot P_{ntb} \cdot R_{ntb}}{P_{ntb} + R_{ntb}} \quad (6)$$

4.2 結果および考察

表 2 に，人手で判定した結果を分類器で学習した際の 10 分割交差検定の結果を示す．NB では 9 割を下回ったが，HNB, WAODE では非常に高い精度で細分化ブロックを正確にタイトルブロック，もしくは他ブロックに分類できている．特徴量間に依存関係が存在するにも関わらず，NB が特徴量間の独立性を仮定しており，事後確率の計算結果が予期しないものとなるためであると考えられる．

他ブロックをタイトルブロックと誤判定した場合，本来であれば上に隣接するブロックと結合されるべきであるが，図 1 の (1) に示すように，結合されずに別々のコンテンツブロックとして分割されてしまう．逆に，タイトルブロックを他ブロックと誤判定した場合，本来であれば上に隣接するブロックとは別のコンテンツブロックとして分割されるべきであるが，図 1 の (2) に示すように，分割されずに一つのコンテンツブロックとして結合されてしまう．Web の閲覧者は，図 1(1) のように分割すべきではないところで細かいブロックに分割されるよりかは，図 1(2) のように 1 つの大きなブロックとして分割されることを好む傾向にある．そのような分割を行うためには，他ブロックをタイトルブロックとして誤判定する数を減らすことが重要である．つまり，タイトルブロックの判定精度 P_{tb}，および他ブロックの再現率 R_{ntb} が高くなるようなアルゴリズムを採用すればよい．HNB アルゴリズムによって生成された分類器は，タイトルブロックの判定精度 P_{tb} が 0.954，他ブロックの再現率 R_{ntb} が 0.958 であり，今回の実験では最も高い性能を有している．以上の理由により，Web ページ分割のためのタイトルブロック抽出には，HNB アルゴリズムが適している．

判定に失敗したブロックの中には，style によってフォントサイズを変更してタイトルを表しているブロックが存在した．提案手法では，タイトルブロックは H1,H2,H3,H4,H5,H6,DT タグを用いて記述されることが多いというヒューリスティクスに基づきタイトルブロックの判定を行っている．しかし，Web ページ製作者によっては style によってフォントサイズを変更したり，また font タグの size 属性を用いて直下に存在するブロックよりもフォントサイズを大きくすることによってタイトルを表現する場合も存在する．したがって，下に隣接するブロックのフォントサイズとの大小関係という特徴量を考慮することにより，精度・再現率が改善する可能性がある．

表 2: タイトルブロックの判定精度と再現率

	NB	HNB	WAODE
a: タイトルブロックを正しく判定した数	801	770	792
b: タイトルブロック以外のブロックを正しく判定した数	733	837	818
c: タイトルブロックを他ブロックと判定した数	35	66	44
d: 他ブロックをタイトルブロックと判定した数	141	37	56
P_{tb}: タイトルブロックの判定精度	0.850	0.954	0.934
R_{tb}: タイトルブロックの再現率	0.958	0.921	0.947
F_{tb}: タイトルブロックの F 尺度	0.901	0.937	0.941
P_{ntb}: 他ブロックの判定精度	0.954	0.927	0.949
R_{ntb}: 他ブロックの再現率	0.839	0.958	0.936
F_{ntb}: 他ブロックの F 尺度	0.893	0.942	0.942

5. 応用

本手法は Web ページからのメインコンテンツ抽出システムへの応用が可能である．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. Web ページ中央（センター）のテンプレートブロックに存在する
2. 情報を伝える役割を持つ
3. ページ上部に配置されている

の 3 つを挙げる．本稿では，ヘッダー，フッター，レフトバー，ライトバー，センターの，5 種類のテンプレートブロックの組み合わせによって Web ページが構成されている前提で Web ページ分割手法を提案した．ヘッダーやフッターに主情報がある可能性は低く，メインコンテンツはセンターのテンプレートブロックに存在するはずである．次に，メインコンテンツは Web ページの閲覧者に対して情報を伝達する役割を持つため，メインコンテンツには長いテキストノードや解像度の高い画像が含まれることが多い．したがって Web ページにおけるメインコンテンツが締める面積の割合は大きくなる傾向にある．最後に，Web ページは縦長のものが多いが，そのような Web ページにおいてメインコンテンツを Web ページ下部に配置してしまうと，ユーザがその情報に辿り着くためには Web ブラウザ上で Web ページをスクロールする必要が発生する．以上の理由により，ユーザビリティを考慮して制作された Web ページは，メインコンテンツをページ上部に配置するはずである．これらのヒューリスティクスを定式化し，Web ページからメインコンテンツを自動抽出するシステムの実装・評価は本研究の今後の課題である．

6. おわりに

本稿では確率モデルに基づいた Web ブロックの役割推定手法を行った．本手法ではまず，Web ページをテンプレートブロックへと分割した後，細分化ブロックと呼ばれる単位まで分割を行う．細分化ブロックの中でも直下の Web コンテンツの内容を説明するタイトルブロックに着目し，コンテンツブロックへの結合を行うことにより Web ページを意味的にまとまりのある単位へと分割を行う．計算機によるタイトルブロックの自動抽出を行うために，タイトルブロックが持つ 9 つのパラメータに着目し，機械学習による分類器を生成した．評価実験により，Naive Bayes(NB)，Hidden Naive Bayes(HNB)，Weightily averaged one-dependence estimators(WAODE) アルゴリズムによって生成した分類器の性能を示した．HNB アルゴリズムによって生成された分類器は，タイトルブロックの判定精度，他ブロックの再現率ともに，最も高い値を示した．Web ページ分割のためのタイトルブロック抽出には，HNB アルゴリズムが適していることが分かった．

参考文献

[1] Y. Lan, L. Bing and L. Xiaoli, “Eliminating noisy information in Web pages for data mining,” Proceedings of the ninth ACM SIGKDD international conference on Knowledge(KDD’03) discovery and data mining, pp.296-305, 2003.

[2] Gen Hattori, Keiichiro Hoashi, Kazunori Matsumoto, and Fumiaki Sugaya: Robust web page segmentation for mobile terminal using content-distances and page layout information, Proceedings of the 16th international conference on World Wide Web(WWW’07), pp.361-370, 2007.

[3] Hui Guo, Jalal Mahmud, Yevgen Borodin, Amanda Stent, and I.V. Ramakrishnan: A General Approach for Partitioning Web Page Content Based on Geometric and Style Information, Proceedings of the Ninth International Conference on Document Analysis and Recognition(ICDAR ’07), pp.929-933, 2007.

[4] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma: VIPS: a Vision-based Page Segmentation Algorithm, Microsoft Technical Report, MSR-TR-2003-79, 2003.

[5] Shumeet Baluja: Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework, Proceedings of the 15th international conference on World Wide Web(WWW’06), pp.33-42, 2006.

[6] Taiki Ito, Hiroyuki Sano, Tadachika Ozono, and Toramatsu Shintani: A Hierarchical Web Page Segmentation Algorithm using Machine Learning, The Eleventh IASTED International Conference on Intelligent Systems and Control(ISC 2008) 2008.

[7] 佐野博之，土井達也，白松俊，大囿忠親，新谷虎松，“役割に基づく Web ページの分割手法とその応用について,” 電子情報通信学会技術研究報告. AI, 人工知能と知識処理 110(301), pp.61-66, 2010.

[8] Jiang Liangxiao, Zhang Harry, and Cai Zhihua, “A Novel Bayes Model: Hidden Naive Bayes,” Proceedings of Canadian Artificial Intelligence Conference, pp.432-441, 2005.

[9] Jiang Liangxiao, and Zhang Harry, “Weightily averaged one-dependence estimators,” Proceedings of the 9th Pacific Rim international conference on Artificial intelligence(PRICAI’06), pp.970-974, 2006.