教師情報を必要としない Webページ群のコンテンツ自動抽出ツールの提案

吉田　光男†　　山本　幹雄††

† 筑波大学第三学群情報学類　〒 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
E-mail: †m.yoshida@mibel.cs.tsukuba.ac.jp, ††myama@cs.tsukuba.ac.jp

あらまし　近年の CMS の普及により、Web ページにメニューや著作権表示などが過剰に付加され、ページに占めるコンテンツ（主要部分）は縮小している。Web ページのコンテンツを抽出することができれば、Web 検索システム、携帯電話向けの Web ページ変換システム、コンテンツフィルタリングシステムなどの精度向上、また、Web ページを利用する研究を促すことが期待できる。本論文では、事前に教師データを準備する必要のないシンプルなアルゴリズムで Web ページ群からコンテンツを抽出する手法を提案する。提案手法は、Web ページをブロック（コンテンツ及び不要部分の最小単位）の集合であると考え、ある特定のページにのみ出現するブロックはコンテンツであるというシンプルなアイデアが基になっている。また、本研究のアルゴリズムを実装したソフトウェアを用いて、Web 上に存在するニュースページからコンテンツを抽出した実験結果について報告する。

キーワード　コンテンツ抽出, 教師無し, 半構造データ, HTML, Web とインターネット, Web サイエンス, データマイニング

Primary Content Extraction from Web Pages without Training Data

Mitsuo YOSHIDA† and Mikio YAMAMOTO††

† College of Information Sciences, and
†† Graduate School of Systems and Information Engineering, University of Tsukuba,
Tennodai 1-1-1, Tsukuba, Ibaraki, 305-8573 JAPAN
E-mail: †m.yoshida@mibel.cs.tsukuba.ac.jp, ††myama@cs.tsukuba.ac.jp

Abstract　In recent years, the proportion of primary content in a Web page has been decreasing as content management systems (CMS's) continue to spread, because CMS's automatically and excessively add unnecessary parts such as menus, copyright displays and so on into the Web page. In this paper, we propose a simple and training data-less method extracting the primary content from a collection of Web pages. We regard a Web page as a set of blocks (minimum unit of primary or non-primary content), and assume that blocks of the primary content are unique and those of non-primary content aren't. We describe experimental results to show performance of the method using real Web pages of the news sites in Japanese and English.

Key words　Primary Content Extraction, Unsupervised, Semi-structured Data, HTML, Web and Internet, Web Science, Data mining

1. はじめに

インターネットが普及した今日、様々な利用者が Web ページを作成し、インターネット上には大量の情報があふれている。2008 年 7 月末に発表された Google のデータによれば、1998 年には 2600 万ページであった Web ページ数は、現在、1 兆ページまで急増している [1]。近年の Web ページの増加は、CMS（Content Management System）(注1)の普及に一因がある。CMS は、設定したページテンプレートに基づき Web ページを生成するため、誰でも簡単に大量のページを作成することができる。すなわち、簡単に大量の情報を発信できるようになったが、反面、各 Web ページにメニューや著作権表示が過

(注1)：Web ページのコンテンツを総合的に管理するシステム

剰に付加されるようになり、ページに占めるコンテンツは縮小している。たとえば、図 1(注2)の Web ページでは、ヘッダ、メニュー、広告、関連記事リストなど不要部分が多々存在することによりページに占めるコンテンツ（破線部分）の割合が低いことがわかる。Web ページのコンテンツを抽出することができれば、Web 検索システム、携帯電話向けの Web ページ変換システム、コンテンツフィルタリングシステムなどの精度向上、また、Web ページを利用する研究を促すことが期待できる。

図 1　Web ページに占めるコンテンツの例

Web ページのコンテンツを調べたところ、ある Web ページのコンテンツは他の Web ページに出現しない傾向があることがわかった。そのため、Web ページのコンテンツ及び不要部分の最小単位（ブロック）を適切に決定することができれば、他のページに出現しないブロックを抽出することにより、コンテンツ抽出が可能になると考えられる。提案手法では、ブロックレベル要素を基にコンテンツ及び不要部分の最小単位である『ブロック』を抽出し、そのブロックが他のページにも出現するか否かを調べることにより Web ページのコンテンツを抽出する。

2. 関 連 研 究

Web ページからコンテンツを抽出する手法は、近年、多くの提案が行われている。Bing ら [2] は、Web ページを一連のセルと見なし、各セルに文字数、句読点数などに応じたスコアを付加した後、そのスコアの大きさを山に見立て、平均的な Web ページではどの山がコンテンツであるかを学習し抽出する手法を提案している。また、鶴田ら [3] は、平均的な Web ページにおいて、ウィンドウのどの位置に主要 DOM ノードが出現するかという情報を用いて主要 DOM ノードを抽出し、その主要 DOM ノードの中からヒューリスティックスで不要部分を除去することによりコンテンツを特定する手法を提案している。これらの手法では、事前に教師データを準備する必要があるほか、平均的な Web ページの構造が変わると抽出が困難になるという問題を抱えている。本研究では、Web ページ群を対象とすることで教師データを必要としない手法を提案し、また、平均的な Web ページの構造に依存しない手法を提案する。

Lin ら [4] は、同じサイト内の Web ページを収集し、ページ中の部分の情報量を計算することによりコンテンツの抽出を試みている。この手法では、計算量が大きくなる傾向があり、Debnath ら [5] は、計算量を小さくした IBDF（Inverse Block Document Frequency）と呼ばれるサイト内におけるページ中の部分の重要度スコアを計算する手法を提案している。しかし、彼らの提案手法には、2 つの問題点がある。1 つ目の問題点は、コンテンツ候補となる部分の抽出に tag-set と呼ばれる『コンテンツと不要部分を分断しやすいタグのリスト』情報が必要であり、この情報は Web ページデザインの流行に左右されることである。彼らは、各ニュースサイトではテーブルタグ（TABLE）により記事本文と不要部分が分断されているため、優先的に分割するのがよいと主張しているが、筆者の調査では、現在、各ニュースサイトではテーブルタグ（TABLE）により記事本文と不要部分が分断されておらず、この知識が古くなっていることがわかっている。2 つ目の問題点は、IBDF を計算した後、各 Web サイトに適した閾値を決定し、コンテンツを抽出するということである。Web 上には無数の Web サイトが存在しており、全ての Web サイトに適切な閾値を決定することは困難である。本提案では、ブロックの抽出に W3C（World Wide Web Consortium）が定義するブロックレベル要素を利用することにより、新たに『コンテンツと不要部分を分断しやすいタグのリスト』を準備する必要のない手法を提案し、また、あるコンテンツは他の Web ページに出現しないという仮説により、各 Web ページに閾値を設定する必要がない手法を提案する。

3. Web ページ群のコンテンツ抽出

3.1 コンテンツとは

一般的に、Web ページは人間が必要とするコンテンツ（主要部分）と不要部分から成り立っている。ニュースの Web ページを例に取れば、記事タイトルや記事本文はコンテンツであり、メニューや著作権表示は不要部分である。本研究では、ニュースの Web ページのコンテンツを次のように定義し、実験・検討を行った。

（1） 記事タイトル
（2） 記事本文
（3） 記事日時
（4） 著者名
（5） 写真・図
（6） 写真・図の説明文
（7） ニュース配信元の著作権情報

不要部分の例としては、広告、メニュー、著作権情報が挙げ

(注2)：http://www.asahi.com/business/update/0106/TKY200901060314.html

られる。広告は、表示されている Web ページとの関連性が高ければコンテンツになりうるが、広告除去ソフトウェア(注3)が存在するなど一般的にコンテンツと認知されていない。また、メニューは、別のページに移動するための情報であり、その表示されている Web ページに必ずしも必要とされていない。そして、著作権情報は、表示されている Web ページが属する Web サイトの情報が記載されており、メニュー同様、その Web ページには必ずしも必要とされていない。ただし、ニュース配信元の著作権情報は、表示されている Web ページのコンテンツそのものの権利情報を表しているため、著者名と同列に扱い、コンテンツとして認めている。

3.2 コンテンツ抽出手法の概要

Web ページのコンテンツを調べたところ、ある Web ページのコンテンツは他の Web ページに出現しない傾向があることがわかった。すなわち、不要部分は複数の Web ページをまたいで何度も出現するが、コンテンツは 1 つの Web ページにのみ出現する傾向がある。したがって、何度も出現する不要部分を除外し、他の Web ページには出現しない部分を抽出すれば、コンテンツを抽出できる。

本研究で提案する Web ページ群のコンテンツ自動抽出手法は、次の 4 つの過程から構成される。

Step.1 [Web ページ群の収集]
コンテンツの抽出を行う Web ページ群を収集する。

Step.2 [ブロックの抽出]
各 Web ページのコンテンツの最小単位となるブロックを抽出する。

Step.3 [特徴ベクトルの生成]
Step.2 で生成した各ブロックの特徴ベクトルを生成する。

Step.4 [ブロック間の比較]
Step.3 の特徴ベクトルに基づき、あるブロックが他の Web ページにも出現するかどうかを調べる。

Step.5 [コンテンツの特定]
Step.4 の比較結果のうち、他の Web ページに出現しないブロックをコンテンツとして抽出する。

3.3 Web ページ群の収集

本研究では、ある Web ページのコンテンツは他の Web ページに出現しないという特定の構造に依存しない仮説を立てた。これにより、Web ページの集合を与えさえすれば、抽出ルールや閾値を必要とせずにコンテンツを抽出する手法を検討する。

本論文では、Web ページ群を $\boldsymbol{S}$ として次のように表現する。

$$\boldsymbol{S} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3, \ldots, \boldsymbol{D}_N\}$$

$\boldsymbol{D}_i (1 \leqq i \leqq N)$ は各 Web ページを指す。

3.4 ブロックの抽出

コンテンツを抽出するためには、コンテンツの最小単位を決定する必要がある。本論文では、コンテンツの最小単位を『ブロック』と呼ぶ。ブロックは、コンテンツの最小単位であるとともに、不要部分の最小単位でもある。Web ページは、マークアップ言語を定義するための言語の一種である SGML（Standard Generalized Markup Language）に基づく HTML タグで記述されているため、その構造は DOM ツリーで表現することができる。ここでは、DOM ツリーからブロックを抽出する方法を説明する。

たとえば、Web ページは図 2 のような HTML コードで記述されている。図 3 は、図 2 を DOM ツリーとして表現（ただし属性情報と改行のみのテキストは省略している）した結果である。DOM ツリーとは、図 3 のように、タグ、テキストなどを葉とする木構造である。

Web ページのレイアウト方法の特徴及び流行を考慮せず、より汎用的にコンテンツ抽出を行う手法を実現するためには、ブロックの抽出も汎用的である必要がある。Web ページで利用されている HTML タグは、WWW で使われる技術の標準化を進める国際団体である W3C（World Wide Web Consortium）によって定められており、この定義に従うことで汎用的な抽出が可能になる。W3C の定めた HTML タグは、Web ページ内の見出し、段落など文書の基本構造を構成するためのブロックレベル要素（H1, P, DIV, TABLE など）と、特定の語の修飾、ハイパーリンクを設置するためのインライン要素（FONT, STRONG, A など）に大分することができる [6]。

本手法では、ブロックの抽出にブロックレベル要素を用いる。ブロックレベル要素を用いてブロックを抽出する際、ブロックがコンテンツや不要部分の最小単位となるように、下位ノードにブロック要素が存在しないように抽出する。

```
<body>
  <div>
    <p>Text 1</p>
    <img src="#" alt="img-alt text">
  </div>
  <div>
    <img src="#" alt="img-alt text">
    <img src="#" alt="img-alt text">
  </div>
  <div>
    <a href="#" title="a-title text">Text 2</a>
    <script>Code</script>
  </div>
</body>
```

図 2 HTML コードの例

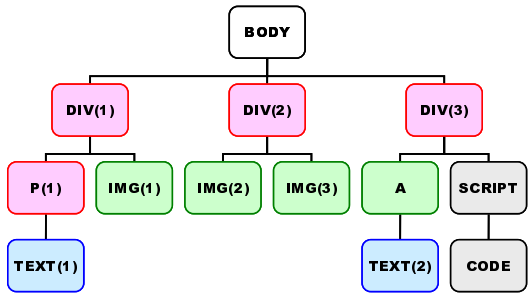

図 3 図 2 の HTML コードを DOM ノードで表現した結果

図 3 の DOM ツリーからブロックを抽出すると、図 4 の通り 5 つのブロックが抽出される。なお、ブラウザにレンダリングさ

(注3)：Adblock (Firefox Add-ons)

れない SCRIPT, STYLE の 2 タグ及びその下位ノードは、ブロック内に含めない。また、BODY タグはブロックレベル要素ではないが、直下にブロックレベル要素以外が存在する HTML 構造にも対応するため、例外的にブロックとして認める。

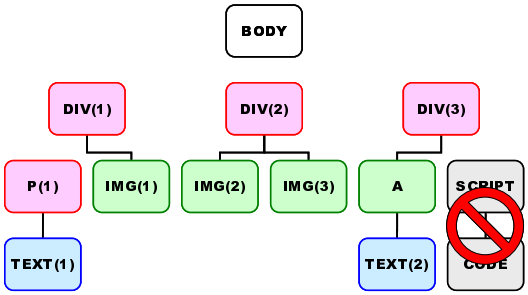

図 4　図 3 の DOM ツリーから 5 つのブロックを抽出した結果

本論文では、Web ページ群 $\boldsymbol{S}$ に含まれる Web ページ $\boldsymbol{D}_i(1 \leqq i \leqq N)$ を次のように表現する。

$$\boldsymbol{D}_i = \{\boldsymbol{B}_{i1}, \boldsymbol{B}_{i2}, \boldsymbol{B}_{i3}, \ldots, \boldsymbol{B}_{iM_i}\} \quad (1 \leqq i \leqq N)$$

$\boldsymbol{B}_{ij}(1 \leqq i \leqq N, 1 \leqq j \leqq M_i)$ は各ブロックを指す。ブロック数は、Web ページごとに変化するが有限である。

3.5　特徴ベクトルの生成

コンテンツの抽出を行うためには、ブロック間の比較が必要になるため、各ブロックの特徴ベクトルを決定する必要がある。本手法では、各ブロックの特徴ベクトル素性として次を用いる。

（1）　ブロック内の各タグの数

（2）　各テキスト（小文字に正規化）の数

（3）　属性 title, alt の各テキスト（小文字に正規化）の数

ブロック内の各タグの数をカウントすることにより、各ブロックのレイアウト情報を表現することができる。また、ブロック内の各テキストの数は各ブロックの内容を推定することができ、属性 title, alt の各テキストの数は、画像（IMG）が出現するブロックの内容を表現することができる。なお、各テキストをカウントする際、テキストを改行によって分割した結果を利用し、空白のみのテキストは除外している。図 4 は属性情報が省略されているが、ブロックを抽出した結果を属性情報を省略せず、HTML コードで表現すると図 5 のようになる（左の番号はブロック番号を表す）。図 5 から本節に基づき特徴ベクトルを生成すると、表 1 のようになる。表 1 の「<a>」「<body>」など括弧で囲まれたものはタグを表し、「a-title text」「text 1」など括弧で囲まれていないものはテキストを表している。

```
1. <p>Text 1</p>
2. <div>
     <img src="#" alt="img-alt text">
   </div>
3. <div>
     <img src="#" alt="img-alt text">
     <img src="#" alt="img-alt text">
   </div>
4. <div>
     <a href="#" title="a-title text">Text 2</a>
   </div>
5. <body></body>
```

図 5　図 2 の HTML コードから 5 つのブロックを抽出した結果

本論文では、Web ページ $\boldsymbol{D}_i(1 \leqq i \leqq N)$ に含まれるブロックの特徴ベクトル $\boldsymbol{B}_{ij}(1 \leqq i \leqq N, 1 \leqq j \leqq M_i)$ を次のように表現する。

$$\boldsymbol{B}_{ij} = (b_{ij1}\ b_{ij2}\ b_{ij3}\ \ldots\ b_{ijL}) \quad (1 \leqq i \leqq N, 1 \leqq j \leqq M_i)$$

$b_{ijk}(1 \leqq i \leqq N, 1 \leqq j \leqq M_i, 1 \leqq k \leqq L)$ は特徴ベクトルの各要素を指す。抽出を行う Web ページ群に含まれる Web ページの数は N であり、テキストはその内容ごとに次元が異なるため L は非常に大きな値を取るが、Web ページ群 $\boldsymbol{S}$ を決定した段階で固定化される。

3.6　ブロック間の比較

3.5 節で述べた特徴ベクトルを用いて、各ブロックが他の Web ページに出現するかどうか、各ブロック同士を比較する（図 6）。ブロック同士を比較する際は、特徴ベクトル同士のコサイン類似度を計算する。特徴ベクトル $\boldsymbol{B}_{ij}(1 \leqq i \leqq N, 1 \leqq j \leqq M_i)$ と $\boldsymbol{B}_{kl}(1 \leqq k \leqq N, 1 \leqq l \leqq M_k)$ の類似度 $Sim(\boldsymbol{B}_{ij}, \boldsymbol{B}_{kl})$ は、次のように計算できる。

$$Sim(\boldsymbol{B}_{ij}, \boldsymbol{B}_{kl}) = \frac{\boldsymbol{B}_{ij} \cdot \boldsymbol{B}_{kl}}{\|\boldsymbol{B}_{ij}\| \|\boldsymbol{B}_{kl}\|}$$

ブロック間の $Sim(\boldsymbol{B}_{ij}, \boldsymbol{B}_{nm})$ が 0.9 を越えた時、それらのブロックは同じであると認める。コサイン尺度を用いることにより、レンダリングにほとんど影響を与えない若干の違いを吸収することができる。

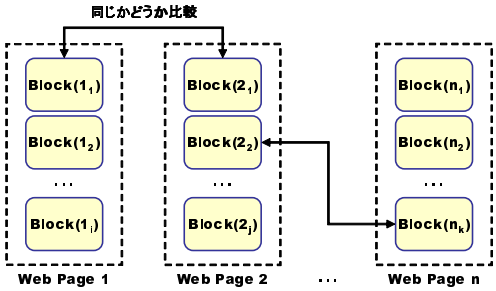

図 6　各ブロックが他の Web ページに出現するかを計算する

3.7　コンテンツの特定

3.6 節で述べたブロック間の比較方法により、他の Web ページには出現しないブロック、すなわち Web ページ群の中で 1 度だけ出現するブロックを抽出する。

4.　実験と考察

4.1　評 価 指 標

本実験の評価尺度は、2. 節で述べた先行研究に倣い、人手で作成した各データセットの適合率（Precision）、再現率（Recall）、F 値（F-measure）を利用したほか、完全一致率（Perfect-matching）というコンテンツを過不足無く認識できた Web ページの割合も利用した。

適合率は、抽出結果として得られたブロック群にどれだけ抽出に適合したブロックを含んでいるかという正確性の指標である。本研究のアルゴリズムによって抽出されたブロックの数を N 、抽出されたコンテンツのうち正解データと適合していたブ

表 1　図 5 を特徴ベクトルに変換した結果

	<a>	<body>		<img>	<p>	a-title text	img-alt text	text 1	text 2
1	0	0	0	0	1	0	0	1	0
2	0	0	1	1	0	0	1	0	0
3	0	0	1	2	0	0	2	0	0
4	1	0	1	0	0	1	0	0	1
5	0	1	0	0	0	0	0	0	0

ロックの数を R とすると、適合率は次のように計算できる。

$$Precision = \frac{R}{N}$$

再現率は、抽出対象としているブロックの中で抽出結果として適合している文書（正解ブロック）のうちでどれだけのブロックを抽出できているかという網羅性の指標である。正解データに含まれるブロックの数を C とすると、再現率は次のように計算できる。

$$Recall = \frac{R}{C}$$

適合率が上がれば再現率が下がり、再現率が上がれば適合率が下がる傾向にあるため、適合率と再現率が用いられる評価には、別途、適合率と再現率の調和平均を取った F 値という尺度がよく用いられる。F 値が高ければ、性能が良いことを意味する。F 値は R を N と C の相加平均で割ったものに相当し、次のように計算できる。

$$\begin{aligned} F\text{-}measure &= \frac{2 \cdot precision \cdot recall}{precision + recall} \\ &= \frac{R}{\frac{1}{2}(N+C)} \end{aligned}$$

本研究では、適合率、再現率、F 値以外に完全一致率というコンテンツを過不足無く認識できた Web ページの割合も利用した。この指標は、一部のコンテンツが各 Web ページで抽出できない場合に低い値を示す。F 値では識別が難しい、別の観点から評価することができる。Web ページ群に含まれる Web ページの数を N 、コンテンツを過不足無く認識できた Web ページの数を M とすると、完全一致率は次のように計算できる。

$$Perfect\text{-}matching = \frac{M}{N}$$

4.2　データセットの作成

実験に使用した正解データは、筆者の開発したアノテーションツール（図 7）を用い、筆者を含め 7 人で作成した。このアノテーションツールは、コンテンツの最小単位となるブロックを自動で認識し、ブラウザ上にてマウスの操作のみで容易にコンテンツをラベル付けすることができる。HTML コードや複雑な DOM ツリーを見ながらラベル付けを行う必要がないため、短時間で、より正確に正解データの作成を行うことができる。

正解データ作成の手順は、まず、筆者が各作業者に 3.1 節と同様に本研究におけるコンテンツとは何なのかを解説した。そして、各作業者がその解説に基づき、それぞれラベル付けを行った。最後に、各作業者がラベル付けした結果を筆者が確認し、必要に応じて修正した上で確定した。この手順を踏むことにより、3.1 節で定義したコンテンツが妥当に定義されているか、すなわち人間がコンテンツの認識を正常に行えるかを確認することができる。

表 2 が示す通り、各作業者の作業の揺れは小さい。従って、3.1 節で定義したコンテンツは妥当に定義されていると考えられる。ごく僅かに揺れているが、この原因は、レンダリングされた『写真・図』と『写真・図の説明文』の区別が難しく一方しか選択されていない、記事日時が経過時間で表記されていたため選択されていないというものが大半であった。なお、複数の作業者によりラベル付けが行われた Web ページが存在するため、表 2 の合計は表 3 と表 6 の合計を超えている。

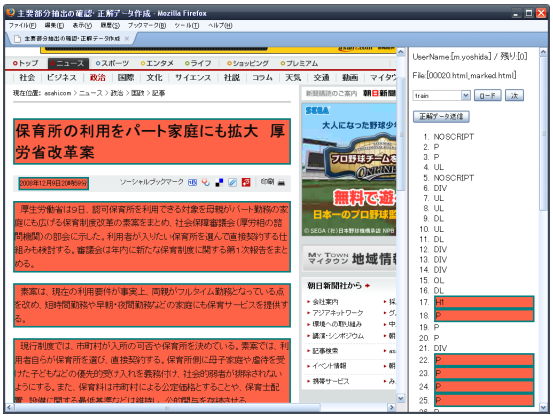

図 7　正解データ作成用のアノテーションツール

4.3　国内のニュースサイトを対象とした実験結果

使用したデータセットの詳細は、表 3 の通りである。asahi.com(注4)、毎日 jp(注5)、YOMIURI ONLINE(注6)の各 Web ページの URL は CEEK.JP NEWS(注7)から取得し、その URL のリストを基に HTML ファイルを取得した。CEEK.JP NEWS から URL を取得する際は、Web ページの内容にばらつきが出ないよう、政治、経済、スポーツのニュースのみにしている。

実験結果を表 4 と表 5 に示す。ALL は、asahi.com、毎日 jp、YOMIURI ONLINE 全てのデータセットを混ぜて実験した結果である。図 8(注8)はコンテンツ自動抽出を行った Web ページの例である（着色部分がコンテンツを示す）。実験結果より、提案手法は全体的に高い性能を示していることがわかる。一方、

(注4)：http://www.asahi.com/
(注5)：http://mainichi.jp/
(注6)：http://www.yomiuri.co.jp/
(注7)：http://news.ceek.jp/
(注8)：http://www.yomiuri.co.jp/politics/news/20081205-OYT1T00914.htm

表 2　人手によるコンテンツ認識の結果

作業者	作業ページ数	適合率	再現率	F 値	完全一致率
筆者	274	0.9968	0.9915	0.9941	0.9526
作業者 A	124	0.9931	0.9558	0.9741	0.6935
作業者 B	69	1.0000	0.9860	0.9930	0.9275
作業者 C	91	1.0000	0.9889	0.9944	0.9560
作業者 D	11	1.0000	1.0000	1.0000	1.0000
作業者 E	43	0.9953	0.9976	0.9965	0.9535
作業者 F	104	0.9977	0.9455	0.9709	0.8173
合計	716	0.9965	0.9771	0.9867	0.8841

表 3　使用したデータセット（国内）

サイト名	ページ数	総ブロック数	正解ブロック数	ページ取得日
asahi.com	179	13593	1031	2008-12-12
毎日 jp	180	28656	1017	2008-12-12
YOMIURI ONLINE	176	33420	1178	2008-12-12
合計	535	75669	3226	-

毎日 jp の再現率が悪く、伴い F 値も悪い結果を示している。また、完全一致率も比較的低い結果を示している。

図 8　実験結果（国内）の Web ページ例（コンテンツ抽出後）

まず、再現率が悪くなる原因を調べたところ、重複した Web ページの存在により、ある Web ページのコンテンツが他のページにも出現することになり、コンテンツとして認識できていなかった。たとえば、図 9(注9)の Web ページと図 10(注10)の Web ページは、別の URL であり右上の不要部分（破線部分）も異なるが、コンテンツは同じである。このような例が毎日 jp データセット内に 18 組存在しており、再現率低下の原因となっている。これを解決するには、ブロック間に比較を行う前に Web ページ間の類似度を計算し、類似性が高い Web ページ間ではブロック間の比較を行わないという方法や、各 Web ページには必ず 1 ブロック以上のコンテンツが存在するという仮説を追加し、コンテンツが 1 ブロック以上抽出されるように、他の Web ページで出現を認める数を自動調整する方法が考えられる。なお、重複した 18 組の Web ページを人手で除外し、実験を行ったところ、適合率 0.9494、再現率 0.9805 を示した。重複した Web ページを検知することにより大幅な性能改善が期待できることがわかる。

図 9　毎日 jp の Web ページ例 1

そして、完全一致率が悪くなる原因を調べたところ、Web ページ内のコンテンツのうち、1 ブロックだけ失敗しているようなケースが多かった。その代表例は、図 11(注11)のような記事日時の日付（破線部分）の抽出である。記事日時に時刻情報

(注9)：http://mainichi.jp/enta/sports/news/20081211k0000e050032000c.html

(注10)：http://mainichi.jp/enta/sports/baseball/news/20081211k0000e050032000c.html

(注11)：http://www.yomiuri.co.jp/atmoney/mnews/20081210-OYT8T00266.htm

表 4　実験結果（国内 1）

サイト名	適合率	再現率	F 値	完全一致率
asahi.com	0.9980	0.9777	0.9878	0.8939
毎日 jp	0.9372	0.7925	0.8588	0.5111
YOMIURI ONLINE	0.9965	0.9559	0.9757	0.8125
合計	0.9800	0.9113	0.9444	0.7383

表 5　実験結果（国内 2）

サイト名	適合率	再現率	F 値	完全一致率
ALL	0.9803	0.9113	0.9446	0.7383

図 10　毎日 jp の Web ページ例 2

図 11　日付の抽出に失敗した例

が含まれない場合、日付の表現方法が限られるため他の Web ページにも出現する可能性が高くなる。これを解決するためには、予め日付の表現方法を学習したモデルを準備し、日付の抽出のみ別途抽出を行うという方法が考えられる。

また、表 4 の合計と表 5 の結果がほぼ同等であるが、抽出方法は異なる。表 4 の合計は、各 Web サイトで Web ページ群を作りコンテンツを抽出した結果の合計であるが、表 5 はデータセット全ての Web ページで 1 つの Web ページ群を作り抽出した結果である。このことから、Web サイトを横断して Web ページ群を作りコンテンツを抽出したとしても、性能にほとんど影響を与えていないことがわかる。

4.4　海外のニュースサイトを対象とした実験結果

使用したデータセットの詳細は、表 6 の通りである。CNN.com(注12)の各 Web ページの URL は Google News（英語版）(注13)から取得し、その URL のリストを基に HTML ファイルを取得した。Google News から URL を取得する際は、ドメインのみを指定し(注14)、Web ページの内容にばらつきが出るようにしている。ただし、閲覧者がコメントを付けられる Blog 形式のページは人手により除外している。

実験結果を表 7 に示す。図 12(注15)はコンテンツ自動抽出を行った Web ページの例である（着色部分がコンテンツを示す）。実験結果より、国内のニュースサイトに比べて比較的悪い結果を示している。特に再現率と完全一致率が悪い結果を示している。

図 12　実験結果（海外）の Web ページ例（コンテンツ抽出後）

CNN.com のデータセットには、毎日 jp データセットと同様

(注12)：http://www.cnn.com/
(注13)：http://news.google.com/
(注14)：検索クエリ「site:cnn.com」を利用した
(注15)：http://sportsillustrated.cnn.com/2009/baseball/mlb/01/15/bp.salarycap/

表 6　使用したデータセット（海外）

サイト名	ページ数	総ブロック数	正解ブロック数	ページ取得日
CNN.com	175	31401	2758	2009-01-16

表 7　実験結果（海外）

サイト名	適合率	再現率	F 値	完全一致率
CNN.com	0.9438	0.7128	0.8122	0.2971

に、別 URL であるがコンテンツが同じという Web ページが 14 組存在おり、これが再現率悪化の主な原因であると考えられる。なお、重複した 14 組の Web ページを人手で除外し、実験を行ったところ、適合率 0.9411、再現率 0.8953 を示した。国内のデータセットの場合と同様、重複した Web ページを検知することにより大幅な性能改善が期待できることがわかる。

また、CNN.com は、著者名とニュース配信元の著作権情報が独立したブロックに記述されているが、これらの情報のバリエーションは他のコンテンツに比べて少ないため、再現率と完全一致率が悪化している。図 13(注16) の Web ページは、ニュース配信元の著作権情報（破線部分）が抽出できなかった例である。

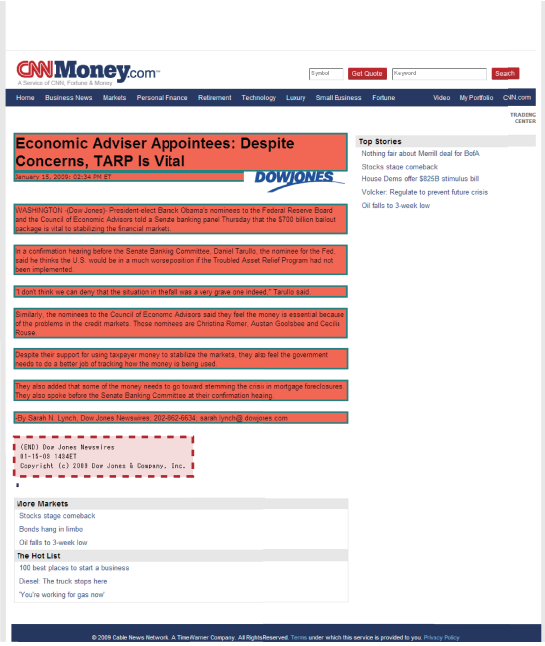

図 13　ニュース配信元の著作権情報の抽出に失敗した例

5. おわりに

本研究では、Web ページの集合を与えさえすれば、抽出ルールや閾値を必要とせずにコンテンツを抽出する手法を検討し、あるコンテンツは他の Web ページに出現しないという仮説を立てた。そして、ブロックレベル要素を基にコンテンツ及び不要部分の最小単位である『ブロック』を抽出し、そのブロックが他のページにも出現するか否かを調べることにより Web ページのコンテンツを抽出する手法を提案した。この提案手法は、一切の教師データを必要としないため、非常に小さな労力で Web ページのコンテンツを抽出することができる。

提案手法を実装したソフトウェアを用い、国内外のニュースサイトを対象に実験を行った。その結果、概ね良好な性能を示したが、再現率は適合率に比べ低い値を示した。再現率の低下は、Web ページ群の中で行われるコンテンツの再利用、日付などバリエーションの少ない情報の抽出失敗が原因だと考えられ、さらなる改良が必要である。

今後は、本研究の成果をソフトウェアとして公開し、Web ページに関する研究の標準的なソフトウェアとなることを目指す。また、本研究による抽出結果を学習データとして利用し、単一の Web ページからコンテンツを抽出する手法の研究を行うことを考えている。

(注16)：http://money.cnn.com/news/newsfeeds/articles/djf500/200901151434DOWJONESDJONLINE001004_FORTUNE5.htm

文　献

[1] Jesse Alpert, Nissan Hajaj. (2008).“ We knew the web was big... ”. Official Google Blog. http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, (Accessed 2009-01-29).

[2] Lidong Bing, Yexin Wang, Yan Zhang, Hui Wang. (2008).“ Primary Content Extraction with Mountain Model ”. IEEE CIT2008. pp.479-484.

[3] 鶴田 雅信, 増山 繁. (2008).“ 未知のサイトに含まれる Web ページからの主要部分抽出手法 ”. 言語処理学会第 14 回年次大会発表論文集.

[4] Shian-Hua Lin, Jan-Ming Ho. (2002).“ Discovering Informative Content Blocks from Web Documents ”. In Proceedings of ACM SIGKDD'02. pp.588-593.

[5] Sandip Debnath, Prasenjit Mitra, Nirmal Pal, and C. Lee Giles. (2005).“ Automatic Identification of Informative Sections of Web Pages ”. IEEE Transactions on Knowledge and Data Engineering. Vol.17, No.9, pp.1233-1246.

[6] W3C. (1999).“ The global structure of an HTML document ”. HTML 4.01 Specification. http://www.w3.org/TR/1999/REC-html401-19991224/struct/global.html#h-7.5.3, (Accessed 2009-01-29).