3F3-5

ウェブページ内の階層構造を考慮した本文抽出技術

Content Extraction from Web Page using Hierarchical Structure

藤田 尚樹 Naoki Fujita　安田 宜仁 Norihito Yasuda　片渕 典史 Norifumi katafuchi　片岡 良治 Ryoji Kataoka

*1 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

Apart from the main content, web pages contain various types of information such as advertisements, menu bars. Identifying the main content of web pages could benefit many applications including text analysis and information retrieval. Conventional methods assume that HTML elements of main content blocks appear continuously. Under this assumption, they label the elements using CRFs. HTML pages have, however, the hierarchical nature and the labels of upper nodes and lower nodes must be consistent. We propose to add an extra step to re-label the result of CRFs using the hierarchical consistency. Experimental results show that our method outperforms the baseline which only considers continuity of blocks.

1. はじめに

インターネット上には大量のウェブページが存在し，多くのページには広告やメニューなど主要コンテンツである本文部分以外の情報も含まれている．それらの本文部分以外の情報はウェブページを解析する際に悪影響を与える．例えば検索エンジンでは本文が検索キーワードと関係無いにも関わらず，検索キーワードが本文部分以外に含まれるために検索結果に表示されてしまうと検索精度が低下してしまう．また，ウェブページ内のテキストを解析して評判分析を行う際にも本文部分以外の情報により精度の低下が生じる．

ウェブページは階層構造を持つ形式で記述され，その階層構造は DOM ツリーとして表すことができる．DOM ツリーではある要素がレンダリング後に表す領域には，その要素の下位要素は全て含まれる．そのため，ある要素が示す領域が本文であれば，その下位要素が示す領域も本文であると言える．我々はこの特性を**階層性**と呼ぶ．また，ウェブページの各要素を深さ優先探索の順で並べた場合には本文領域は連続しやすく，我々はこの特性を**連続性**と呼ぶ．

Marek ら[Marek 2007]はウェブページをテキストのシーケンスとして扱い，CRF[Lafferty 2001]を用いて本文を抽出する手法を提案している．この手法は我々の定義する連続性に着目した手法であると言える．我々は Marek らの手法を階層性に着目して拡張することで，本文抽出精度の向上が可能と考える．

本論文では，ウェブページから本文部分を抽出する手法として，従来手法と同様に連続性に着目した一次判定と，階層性に着目した二次判定を組み合わせた手法を提案する．一次判定ではウェブページを特定のタグ毎にブロックに分割し，そのブロックのシーケンスに対して本文を判定する．本文部分は階層性に対して整合性が取れている必要があるため，二次判定において一次判定結果がブロックの各親子関係で整合性が取れているか多数決モデルにより再判定を行う．

2. 関連研究

ウェブページからの本文抽出手法として Marek らは CRF を用いてウェブページの本文部分を抽出する手法を提案している．Marek らの手法では，まず HTML からテキスト要素のみをブロックとして抽出し，テキストブロックのシーケンスを作成する．次に各ブロックにおける素性をタグ情報やテキストの内容から計算する．そして，計算された素性を用いて CRF による「本文」「本文外」のラベリングを行い，「本文外」と判定された部分を削除して本文部分を抽出する．この手法では，先に挙げたウェブページの 2 つの特性のうち，連続性は考慮しているが，階層性については考慮されていない．我々は階層性を上手く利用することで更なる精度向上が可能と考える．

連絡先：藤田 尚樹 fujita.naoki@lab.ntt.co.jp
日本電信電話株式会社 NTT サイバーソリューション研究所
神奈川県横須賀市光の丘1－1

3. 提案手法

我々は Marek らの手法を拡張し，階層性も考慮した本文抽出手法を提案する．3.1 ではウェブページの特性である階層性と連続性について詳細に述べ，3.2 にその特性を考慮した手法の概要，3.3 にその具体的な手順を記述する．

3.1 階層性と連続性

1 章で述べた階層性・連続性の 2 つの特性について図 1 を用いて詳細に述べる．図中の左は典型的なウェブページのレイアウトを記述している．ページを<DIV>, <P>, <TD>, <BODY>のタグで分割すると，大きく B1～B5 ブロックに分けられ，例えば B1 はヘッダー，B2 はメニューやリンクリストなどが表示される．B1～B5 は配下に B1-1 などの下位ブロックを有しており，図 1 右上のようなツリー構造で表せる．下位ブロックはレンダリング後には上位ブロックの領域内に配置され，B3 が表す領域が本文部分の場合，その配下の B3-1, B3-2, B3-2-1, B3-2-2 及び B-3-3 も本文部分となる．このように上位ブロックが本文である場合に下位ブロックも本文となる特性を階層性と呼ぶ．
また，図 1 のツリー構造を深さ優先探索順に並べると，”B1→B1-1→B1-2→B2→B2-1→…→B5-1” となり，上位ブロックの次には下位ノードが続く．B3 が本文領域であれば，それに含まれるブロックは”…→B3→B3-1→B3-2→B3-2-1→B3-2-2→B3-3→…”のように連続することになり，我々はこれを連続性と呼ぶ．

3.2 手法概要

我々は，Marek らと同様に連続性に着目してブロックのシーケンスから CRF を用いて本文部分を判定する一次判定と，一次判定の結果を階層性に着目してブロックの各親子関係で整合性が取れているか多数決モデルによって再判定を行う二次判定による本文抽出手法を提案する．

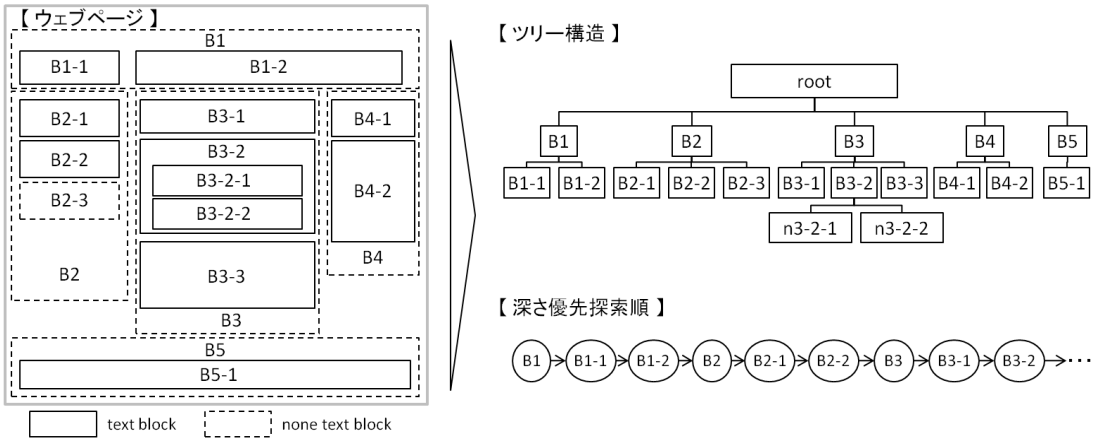

図 1. ウェブページと構成ブロックの関係

(1) 一次判定

HTML 文書から深さ優先探索順でテキストを含むブロックのシーケンス(テキストブロックシーケンス)と，各テキストブロックの親ブロックでテキストブロックシーケンスに含まれないブロックのシーケンス(親ブロックシーケンス)の 2 つのブロックシーケンスを抽出する．各シーケンスに対して各ブロックの素性を抽出し，CRF を用いて一次判定を実施する．各ブロックから抽出する素性は HTML タグ情報及びテキスト情報などから生成する．

(2) 二次判定

一次判定により，全てのテキストブロックとその親ブロックが一次判定結果を有することになる．2 つのブロックシーケンスに含まれる各親子関係において多数決モデルによる二次判定を行う．1 つの親子関係には 1 つの親ブロックと複数の子ブロックが含まれる．その中で，一次判定で本文と判定されたブロックの割合が上位閾値以上ならば，その親子ブロックは全て本文と判定し，下位閾値未満ならば親ブロックを本文外とし子ブロックは変更しない．また，本文の割合が上位閾値と下位閾値の間の場合は判定結果の変更は行わない．上位閾値・下位閾値は事前に設定した 0 から 1 の間の値を用いる．各親子関係の判定の際には，子ブロックは事前にさらに子ブロックと二次判定を実施し，その結果を用いる．

2 つの閾値を用いる理由は子ブロックの最後のみ本文外である場合が多いためである．例えば図 1 の B3 を親ブロックとする親子関係で B3-1 と B3-2 が本文で B3-3 が広告であるような場合である．この場合，親子関係内の本文の割合が高いにも関わらず，親子全てを本文とすると，本文外の情報を抽出してしまう．そのため，親子全てを本文とするための閾値は高く設定したい．しかし，その閾値のみでは閾値以下の場合に親ブロックを本文外とする場合が多くなり，ツリー構造の末尾から二次判定をしていくと大半のブロックが本文外となりがちである．そのため上位・下位の 2 つの閾値を用いる．

二次判定の内容図 2 を用いて具体的に説明する．図中左に HTML 中の一部の親子ブロックを示す．図中中央上部のように一次判定結果が親ブロックと子ブロックの左から 3 つ目のみ本文外と判定された場合，親子関係に含まれるブロックの 6 割が本文と判定される．従って上位閾値が 0.6 以下であれば全ブロックを本文と判定する．また，図中中央下部のように一次判定結果が親ブロックと 2 つ目の子ブロックのみ本文と判定された場合，親子関係に含まれるブロックの 4 割が本文と判定されている．下位閾値が 0.4 以上の場合，親ブロックを本文外とし，子ブロックは変更しない．

二次判定のアルゴリズムを図 3 に示す．判定対象ブロック **b**，一次判定結果 **R1**，二次判定結果 **R2** を入力として **R2** を更新するアルゴリズム **Second_Evaluation** はまず **b** の子ブロック集合 **C(b)**を取得し，各子ブロックが二次判定されているかを確認する．二次判定されていなければ，その子ブロックを判定対象ブロックとして再帰的に **Second_Evaluation** アルゴリズムを実行する．その後，子ブロックの二次判定結果と **b** の一次判定結果から多数決モデルにより，上位閾値(**upper_threthold**)以上の割合が本文であれば，全てを本文とし，逆に本文の割合が下位閾値(**lower_threthold**)未満であれば，**b** は本文外する．また，どちらでもない場合は R1 の値を引き継ぐ．

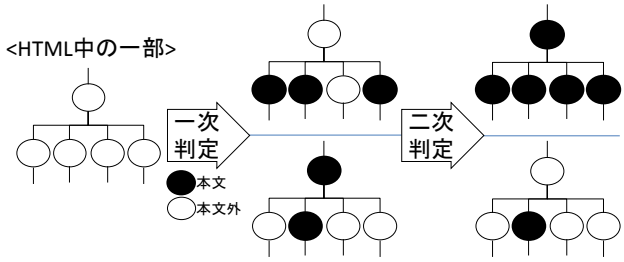

図 2. 二次判定の例

・INPUT
b : target block
R1(b) : block b's result of First - Evaluation (1: 本文, 0 : 本文外)
R2(b) : block b's result of Second - Evaluation (1: 本文, 0 : 本文外)
・OUTPUT
R2(b) : updated block b's results of Second - Evaluation
・ARGORITHM

```
Second_Evaluation{
  b = $1;
  C(b) = ブロックbの子ブロックの集合;
  foreach c in C(b){
    unless( exists(R2(b)){Second_Evaluation(c);}
  }
  ratio = (R1(b) + Σ_{c∈C(b)} R2(c))/(| C(b) | +1);
  if( ratio > upper_threthold){
    R2(b) = 1;
    foreach c in C(b) { R2(c) = 1;}
  }elsif(ratio < lower_threthold){
    R2(b) = 0;
  } else{
    R2(b) = R1(b);
  }
}
```

図 3.二次判定のアルゴリズム

3.3 手順

提案手法による本文を抽出する際の手順を述べる.

(1) ウェブページの整形・形式標準化

ウェブページの記述形式は W3C[1]により複数の標準が規定されているため, XHTML1.0 に統一する. また, コメントアウト部分などを削除する.

(2) シーケンス生成

ウェブページの要素を深さ優先探索順に並べ, <DIV>, <P>, <TD>, <BODY>のタグ毎にブロックとして区切る. その中からテキストを含むブロックのシーケンス(テキストブロックシーケンス)と, その親ブロックでテキストを含まないブロックのシーケンス(親ブロックシーケンス)を生成する.

(3) 素性抽出

シーケンスに含まれる各ブロックから素性を抽出する.

(4) CRF のモデル学習

学習用データを用い CRF のモデル学習を行う. モデルはシーケンス毎に 2 種類作成する.

(5) CRF による一次判定

テキストブロックシーケンスと親ブロックシーケンスに対し(4)で作成した各モデルを用いて CRF による一次判定を行う.

(6) ツリー構造を考慮した二次判定

一次判定結果を用いて 3-2-(2)で述べた二次判定を行う.

(7) 本文抽出

二次判定によって本文と判定されたブロックのテキストを抽出.

4. 評価実験

4.1 概要

提案手法による本文抽出の精度評価実験を実施した. 実験には 4.2 に示す独自データセットを用いた. 二次判定の上位・下位閾値は 0.1～0.9 の間を 0.1 刻みで変化させ, 各値での精度を確認した. 提案手法の有効性を確認するため, Marek らのテキストブロックシーケンスに対して CRF を用いて本文を抽出する手法をベースラインとして提案手法と精度比較を行った.

4.2 データセット

実験には独自でクロールした 7,308 ページを用いた. 収集は全 4,845 サイトから 1 サイト最大 2 ページを収集した. 各ページに対して人手で本文部分の判定を実施し, 正解データを作成した. ウェブページは学習に 5,917 ページ, 評価に 1,391 ページを用いた.

4.3 評価指標

評価の際に判定の正解率を用いることも考えられるが, 実用的には本文部分のテキストが正しく全て抽出できているかが重要である. そのため本論文では要約技術の評価などで用いられる機械翻訳で用いられる BLEU[Papineni 2002]と要約の評価で用いられる ROUGE[Lin 2004]を用いて本文抽出の精度評価を行う. BLEU は適合率に関する指標であり, 抽出したテキストが正しく本文部分であるかを示す. ROUGE は再現率に関する指標で本来の本文をどの程度抽出できたかを示す. 再現率と適合率はトレードオフの関係になりやすく, 一方を高めようとすれば他方が低下しがちである. そのため双方の指標を同時に用いることで, 手法の実用性に着目した評価が可能となる.

(1) BLEU

BLEU は 1-gram から n-gram までの適合率の重み付き和であり, 下記式で定義される. 本実験での n の値は一般的に BLEU で評価する際に用いられる n=4 とした.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log p_n\right)$$

$$p_n = \frac{\sum_{j=1}^{K} \min(s_j^n, r_j^n)}{\sum_{j=1}^{K} s_j^n}$$

式中の s_j^n は本手法で出力した本文に含まれる j 番目の n-gram の出現数を表す. また, r_j^n はそのj番目の n-gram の正解本文中の出現数である. ここで BP はペナルティ値として用いられている. BP は再現率が低い場合にスコアが過度に上昇しないようにするパラメータである. しかし, 出力本文の再現率に関しては ROUGE を用いて評価するため, 本実験では BP=1 とし, BLEU では再現率を考慮せず評価を実施した.

(2) ROUGE

ROUGE は正解本文と本手法で出力した本文を比較して n-gram で再現率を計算する手法であり, 下記式で表される.

$$ROUGE-n = \frac{\sum_{j=1}^{K} \min(s_j^n, r_j^n)}{\sum_{j=1}^{K} r_j^n}$$

式中の分母は正解本文中に含まれる n-gram の総数であり, 分子は本手法で出力した本文と正解本文で一致した n-gram の総数である. 本実験での n の値は簡易的に n=2 を用いた.

4.4 実験手順

3.3 の手順に沿って実験を実施する. (3)素性抽出では下記の項目の素性を約 300 個抽出した.

- タグ情報:ブロックがどのタグで分割されているか
- キーワード:人手で選定した, ページのフッターやメニューなどに局所的に頻出する特徴的な語の出現数. 例えばフッターなどに頻出する「copyright」「推奨環境」など
- テキスト量:自ブロック, 子ブロック, ページ全体毎の量
- 句読点, 改行, スペース, 時間表現の数
- ブロック内のリンクの数:<a>,<link>,<li>の数
- テーブル情報:ブロックがテーブル内の何行目,何列目か. 含まれるテーブルの列数, 行数

また, (4)CRF のモデル学習及び(5)CRF による一次判定では CRF++[2]を用いた.

5. 実験結果及び考察

5.1 実験結果

実験の結果, ベースラインと提案手法は 6 割以上のページで同じ本文を抽出した. そのため手法間の精度差を明確にするため, 抽出本文に差分があったページに限定した精度を記述する. 図 4 にベースラインと比較した際の精度向上率をヒートマップで示す.

[1] http://www.w3.org/

[2] http://crfpp.sourceforge.net/

表 1. 抽出本文に差分のあったページの精度評価結果(括弧内は全てのページで評価した場合の値)

上位閾値, 下位閾値	BLEU 提案手法	BLEU ベースライン	BLEU 変化率	ROUGE 提案手法	ROUGE ベースライン	ROUGE 変化率	ページ数
0.7, 0.4	0.776 (0.827)	0.776 (0.826)	+0.0%	0.946 (0.949)	0.931 (0.943)	+1,6%	618
0.7, 0.5	0.780 (0.827)	0.780 (0.826)	+0.0%	0.941 (0.947)	0.932 (0.943)	+1.0%	639
0.8, 0.4	0.775 (0.828)	0.771 (0.826)	+0.5%	0.939 (0.944)	0.936 (0.943)	+0.3%	521

BLEU は上位閾値及び下位閾値共に高い場合にスコアが上昇した. ROUGE では両閾値共に低い場合にスコアが上昇した. BLEU は最大 2.0%, ROUGE は最大 2.5%精度向上が見られた. この結果から, 両閾値が高くなると抽出するテキストの量は少なくなり, 本文の中で抽出されない部分が増えるが, 本文外の部分は含まれにくくなる. それに対し, 両閾値が低くなると抽出するテキストの量は増加し, 本文の中で抽出される部分が増えるが, 同時に本文外の部分も抽出されやすくなる. 両指標のスコアはトレードオフの関係であるが, 両スコアともベースラインを上回った閾値の組み合わせは 9 組あり, それらは上位閾値が 0.7～0.9, 下位閾値が 0.1～0.5 の範囲の組み合わせであった. その上位 3 組とベースラインのスコアを表 1 に示す. BLEU はベースラインと僅かな差だが, ROUGE では最大 1.6%向上した. この場合, 抽出したテキストが本文である割合は変わらずに, より多くの本文部分が抽出されていることになり, ベースラインより精度が向上したと言える.

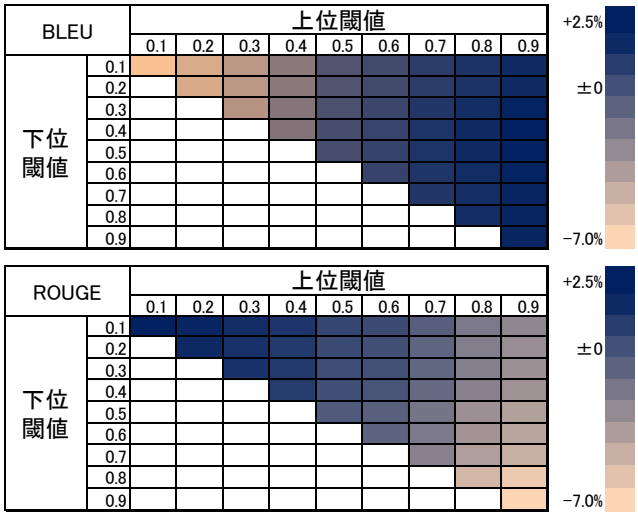

図 4. 各閾値におけるベースラインからの精度変化

5.2 考察(事例分析)

上位閾値を 0.7, 下位閾値を 0.4 とした提案手法とベースラインで精度変化が大きいページを個々に分析した. BLEU に比べ精度変化が大きい ROUGE が 0.4 以上向上もしくは低下した 6 ページを HTML の特徴から 3 つに分類し, それぞれの精度変化を表 2 に示す.

“本文がテーブルの入れ子構造でレイアウトされているページ”では精度向上が見られた. このページは本文を記述する際に<table>タグ内の<td>タグ内で更に<table>タグが複数個使われており, 一部の<td>タグでは短いテキストしか含まれない. 短いテキストを含むブロックは, メニューやリンクリストなど本文外で見られることが多く, CRF では本文外と判定される可能性が高い. しかし, 提案手法の二次判定は同じ<table>タグに含まれる全<td>タグを一つの親子関係に含めて行われる. そのため, 一次判定で本文外と判定された短いテキストを含む<td>タグもテーブル内で本文と判定された<td>タグが多かったため, テーブル全体が本文と二次判定され, 抽出できたと考えられる.

“携帯対応ページ“と“<li>タグで本文の大半が表示されているページ“では精度が低下した. 前者は深い階層構造で記述されることは少なく, 各ブロックは HTML に記述されている通りの順で上から 1 列で表示される. そのため, 単にブロックシーケンスとして扱う方が良いと考えられる. 後者に関しては, タグはページのメニューバーや広告エリアで多数利用されている場合が多いため, 本来は本文である部分が本文外と判定される可能性が高いため精度が低下したと考えられる. このようなページはベースラインでも ROUGE が 0.6 程度と全体と比べて低い精度ということから, 提案手法の一次判定結果がほとんど本文外と判定されてしまい, 多数決モデルの二次判定で多くの親子関係が下位閾値以下となり更に精度を低下させたと考えられる. これに対しては, 素性や学習ページを修正して一次判定で用いる CRF 自体の精度向上が必要である.

表 2. ROUGE が大きく変化したページの特徴と精度変化

特徴	精度変化
本文がテーブルの入れ子構造でレイアウトされているページ	向上
携帯対応ページ	低下
タグで本文の大半が表示されているページ	低下

6. まとめ

本論文ではウェブページの特性である階層性と連続性を考慮した本文抽出手法を提案し, 評価実験の結果, ベースラインに比べて精度向上が見られたことを示すと共に, 提案手法の効果が特徴的なページを個々に分析し, 本文がテーブルの入れ子構造でレイアウトされているページに対して特に有効であることを示した. また, 閾値の調整により再現率と適合率のどちらを重視するかを調整することが可能であることも示した. そのため, 例えば検索エンジンでは再現率を重視するため閾値を低く設定し, 評判分析では適合率を重視するため閾値を高く設定することでそれぞれに適した本文の抽出が可能である. このため提案手法は実用面で有効性が高いと考えられる.

参考文献

[Lafferty 2001] J.Lafferty, A. McCallum, and F. Pereira: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, In Proc. ICML01, 2001.

[Lin 2004] Chin-Yew Lin: ROUGE: A Package for Automatic Evaluation of Summaries, In Proc. ACL04 Workshop on Text Summarization, 2004.

[Marek 2007] Michal Marek, Pavel Pecina, and Miroslav Spousta: Web Page Cleaning with Conditional Random Fields, In Cahiers du Cental, 2007.

[Papineni 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu: BLEU: a Method for Automatic Evaluation of Machine Translation, In Proc. ACL, 2002.