ウェブページにおける非コンテンツ領域の検出

中村　達也　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{tatsuyan, kshirai}@jaist.ac.jp

1 はじめに

本論文はウェブページにおける非コンテンツ領域を自動的に検出する手法について述べる [3]．ウェブページは何らかの情報を発信していると考えられるが，非コンテンツ領域とはその主たる内容を含まず，したがって特に有用な情報も含まない領域と定義する．非コンテンツ領域の典型的な例は広告，ナビゲーションのための目次やツールバー，著作権表示などである．このような非コンテンツ領域を自動的に検出することができれば多くのウェブ情報処理の場面で有益である．例えば，情報検索の場合，非コンテンツ領域内の単語を検索インデックスから除くことにより情報検索の速度や精度の向上が期待できる．一方，ウェブマイニングの場合でも，非コンテンツ領域を自動的に検出しそれをマイニングの対象からはずすことによって処理時間を短縮できる．本論文では，非コンテンツ領域を示唆するキーワードやDOM ツリーから得られる HTML タグなどを素性とし，非コンテンツ領域を検出するモデルを自動的に学習する手法を提案する．

非コンテンツ領域の検出に関する先行研究としては Lin らの研究がある [2]．Lin らは，ニュースサイトを対象に，記事本文の領域とそれ以外の領域を区別する手法を提案している．これに対し本研究では，ニュースサイトに限らず一般のウェブページを対象とした非コンテンツ領域の検出を試みる．また，非コンテンツ領域の検出は，ウェブページにおける意味的なまとまりを検出することを含むという点でウェブページの構造解析に関する研究 [1, 4, 5] と関連が深い．本研究では，ウェブページの階層的な構造を求めることはしないが，DOM ツリーの利用などウェブページの構造解析によく用いられる手法も取り入れている．

2 提案手法

2.1 非コンテンツ領域

1 節で述べたように，ウェブページにおける非コンテンツ領域とは，ページの主たる内容と比べて有用でない情報しか含まない領域である．例えば，図 1 の毎日新聞社のサイト[1]のページは，新聞記事の本文が主たる内容であると考えられる．ここで図 1 中の破線で囲まれた部分に着目する．A,B はナビゲーション目的のリンクであり，C は広告である．これらの領域は新聞記事の本文に比べて有用な情報であるとは言い難い．したがって，このような領域を自動的に検出することができれば，多くのウェブアプリケーションに対して処理時間の短縮やパフォーマンスの向上が期待できる．

本研究では，非コンテンツ領域の検出を様々なウェブアプリケーションの前処理と位置付けている．ところが，非コンテンツ領域の定義は後続するウェブアプリケーションによって異なると考えられる．例えば，ウェブ検索エンジンにとっては，図 1 における A や B の領域は非コンテンツ領域であると考えられる．あるクエリに対し，この領域に含まれるキーワード (例えば「コラム」) にヒットして図 1 のページが取り出されたとしても，このページが適合文書である可能性は低いと考えられるからである．したがって，A,B のような領域は非コンテンツ領域とし，この領域中の単語はページの索引語としない方がよい．一方，ウェブにおけるリンク構造を解析するときは，A や B のようなナビゲーション目的のリンクは重要であるため，これらの領域は当然非コンテンツ領域とするべきではない．

このような状況を考慮し，本研究では正解付きデータから非コンテンツ領域を自動的に学習することを試みる．上述のように非コンテンツ領域の定義はアプリケーションによって異なるが，それに応じた正解データを用意することにより，様々なウェブアプリケーションにある程度柔軟に対応できる．ただし，学習に有効な素性が非コンテンツ領域の定義によって異なることは十分考えられ

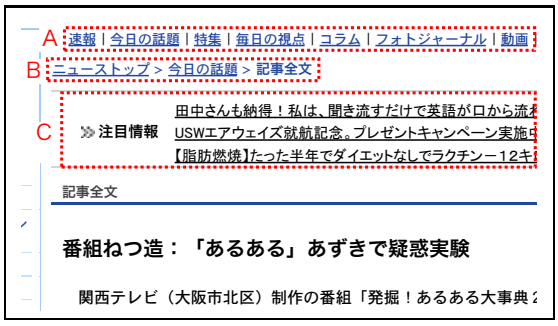

図 1: 非コンテンツ領域の例

[1] http://www.mainichi-msn.co.jp/

る．そのため，あるウェブアプリケーションのために学習した非コンテンツ領域検出モデルを別のアプリケーションに適用した場合，同程度の精度で非コンテンツ領域が検出できるわけではない．しかしながら，ルールベースの検出手法に比べればある程度のポータビリティを持つと考えられる．

なお，本論文では後続のウェブアプリケーションとして情報検索を想定する．すなわち，ある領域に含まれる単語が情報検索の索引語として有効であるかどうかによって非コンテンツ領域か否かを判定する．具体的には以下のような領域を非コンテンツ領域と定義する．

- 広告
- アクセスカウンタ
- ナビゲーションを目的としたリンク
  ページの目次，サイトマップ，リンクのグループなどが該当する．ただし，リンクは内部リンク (同一サイトへのリンク) に限定し，いわゆる外部リンク (他のサイトへのリンク) から構成されるリンク集などは該当しないとする．
- 著作権表示
  「このページの著作権は×××に属します」といった記述．
- 検索フォーム
- 印刷のナビゲーション表示

2.2 非コンテンツ領域の検出

本項では非コンテンツ領域を検出する手法について述べる．まず，対象ウェブページをテキストユニット (Text Unit;TU) に分割する．ここでテキストユニットとは，HTML タグで自動的に分割されたテキストの断片を指す．TU の例を図 2 に示す．図 2 (b) は (a) のウェブページに対応する HTML のソースであり，HTML タグを除く各行が 1 つの TU に相当する．

非コンテンツ領域は一般に複数の TU で構成される．そのため，本研究では，非コンテンツ領域の検出を TU に対するチャンキング問題と捉える．すなわち，ページ中の全ての TU に対して以下の B,I,O のいずれに該当するかを判定する．

B: 非コンテンツ領域の先頭に該当する TU

I: 非コンテンツ領域の先頭以外に該当する TU

O: 非コンテンツ領域ではない TU

例えば，図 2 (a) のページにおいて，ページ上部の内部リンク (「トップ プログラム 会場」の部分) が非コンテンツ領域であるとき，各 TU に対して割り当てられるべきラベルを図 2 (b) の矢印の右に示す．

(a)

トップ　プログラム　会場
大会プログラム
1 日目
...

(b)

```
<table><tr><td><a href="index.html">
トップ                                    → B
</a></td></tr><tr><td><a href="p.html">
プログラム                                → I
</a></td></tr><tr><td><a href="v.html">
会場                                      → I
</a></td></tr></table><h1>
大会プログラム                            → O
</h1>
1 日目                                    → O
```

図 2: テキストユニットと BIO ラベル

チャンキングモデルは教師あり学習により獲得する．すなわち，正しい非コンテンツ領域が付与されたウェブページ集合から TU のラベルを判定するモデルを学習する．学習にはチャンカーツール YamCha[2]を利用した．YamCha は学習アルゴリズムとして Support Vector Machine を採用した汎用チャンカーである．

2.3 素性

非コンテンツ領域検出モデルの学習に用いる素性を以下に述べる．

- 非コンテンツ領域を示唆するキーワード
  「広告」「検索」「著作権」など，非コンテンツ領域を示唆するキーワードをいくつか選別し，それらのキーワードが TU に含まれるか否かを素性とした．キーワードの選別方法は 2.3.1 で述べる．
- テキスト長
  TU の文字数が少なければ少ないほど非コンテンツ領域になりやすいと考えられる．ここでは TU の文字数を 1,2,3-5,6-8,9-15,16 以上のいずれかに分類し，素性として用いた．
- TU が動詞を含むか，形容詞を含むか
  TU が動詞や形容詞といった用言を含む場合は，文または文章が TU に含まれ，コンテンツ領域である可能性が高いと考えられる．
- TU が内部リンク/外部リンク/リンクではないか
  TU が内部リンクのアンカーテキストであるときは非コンテンツ領域である可能性が高いと考えられる．なお，TU を囲む a タグの参照 URL が相対パスのとき，あるいは参照 URL のホスト名が対象ウェブ

[2] http://chasen.org/%7Etaku/software/yamcha/

ページと同一のときには内部リンクとし，それ以外は外部リンクとみなした．

- DOM ツリーにおける HTML タグ
  HTML タグは非コンテンツ領域の検出に有力な手がかりになると考えられる．ここでは，対象ウェブページの DOM ツリーにおいて，TU からルートへ DOM ツリーを辿ったときに到達する 1 番目，2 番目，3 番目の HTML タグ名を素性とした．ただし，font のような文字装飾タグがあったときはそれを無視し，さらに上を辿って素性とする HTML タグを得た[3]．
- DOM ツリーにおける深さの変化
  直前の TU に比べたときの DOM ツリーにおける TU の深さの変化 (同じ，浅くなる，深くなる) を素性とした．これは，DOM ツリー上で深さが変化するときは HTML タグの構造も大きく変わり，非コンテンツ領域とコンテンツ領域の境界になりやすいと考えたためである．
- table 内のリンクの割合
  TU が table タグ内に含まれるとき，式 (1) の値を素性とした．

$$\frac{a\ \text{タグで囲まれた}\ TU\ \text{の数}}{\text{table 内の}\ TU\ \text{の総数}} \quad (1)$$

  また，同じ table 内の TU に対するこの素性は全て同じ値となる．

  ナビゲーション目的のリンクのような非コンテンツ領域は 1 つの table で構成されていることが多いが，その table の中に例外的にリンクではない TU があると，その TU だけはコンテンツ領域と誤って判定されやすい．このような誤りを回避するために導入した素性である．
- table 内のテキストの平均長
  TU が table タグ内に含まれるとき，その table における全ての TU の平均長を素性とした．ただし，値は l_a=0, l_a=1, $1 < l_a < 4$, $l_a \geq 4$ のいずれかとした (l_a は実際の平均長)．この素性を導入した理由は前の素性とほぼ同じである．

2.3.1 非コンテンツ領域を示唆するキーワードの選別

前述のように，本研究では非コンテンツ領域を示唆するキーワードの有無を素性として用いる。キーワードは，訓練データに含まれる全ての名詞 w のうち，以下の 3 つの条件を満たす名詞とする．

- 出現頻度が 20 以上である．
- 式 (2) の値が 0.7 以上である．

$$P_w = \frac{w\ \text{が非コンテンツ領域に出現する回数}}{w\ \text{の出現回数}} \quad (2)$$

- 式 (3) の値が 2 以上である．

$$S_w = P_w \times D_w \quad (3)$$

ここで，P_w は w が非コンテンツ領域に出現する確率である．また，D_w は w が非コンテンツ領域に出現するウェブページのドメインの異り数であり，キーワード w の汎用性を考慮した指標である．例えば，あるキーワードが非コンテンツ領域に頻繁に現われ，P_w の値も高いが，そのキーワードが非コンテンツ領域に出現するのはあるドメインのウェブページのみとする．このとき，そのキーワードはあるウェブサイトの非コンテンツ領域によく出現するだけであり，非コンテンツ領域を示唆する一般的なキーワードであるとは言えない．D_w の値が大きいということは，そのキーワードが様々なウェブサイトの非コンテンツ領域に使われているということであり，非コンテンツ領域を示唆する一般的なキーワードである可能性が高いと考えられる．

3.1 項で述べる実験用コーパスから非コンテンツ領域を示唆するキーワードを抽出したところ，47 個のキーワードを得た．そのほとんどが非コンテンツ領域を示唆するキーワードとして妥当であった．例を以下に示す．

C, Co, ホーム, Copyright, TOP, All, トップ, マップ, Reserved, HOME, プライバシー

3 実験

3.1 実験データ

WWW から実験に用いるウェブページをランダムに収集した．具体的には，まず Open Directory プロジェクト dmoz[4] のウェブディレクトリからランダムに 46 ページを選択した．これら 46 ページならびにこれらのページからリンクを 1 回辿って得られるページを収集した．ただし，frame タグを使っているページは今回の実験の対象外とした．最終的に 781 のウェブページを収集した．

これらのウェブページに対して非コンテンツ領域を人手で付与した．作業は著者 1 名を含む大学院生 4 名で行った．2.1 項で述べた非コンテンツ領域の定義を作業者に説明し，それに従ってページ内の非コンテンツ領域をマークアップさせた．

非コンテンツ領域の判定が異なる作業者間でどの程度一致するかを調べるために，62 ページについては 2 名の作業者に非コンテンツ領域の付与を依頼し，その結果

[3] 具体的には次のタグを無視した．div, font, a, span, strong, select, option, pre, small, kbd, b.

[4] http://dmoz.org/World/Japanese/

表 1: 作業者による非コンテンツ領域判定の一致度

作業者	ページ数	一致度	
		(領域単位)	(TU 単位)
T_1-T_2	36	0.58	0.81
T_3-T_4	26	0.64	0.84

を比較した．表 1 は 2 人の作業者によって付与された非コンテンツ領域の一致度を示している．一致度の定義を式 (4) に示す．

$$\frac{2 \times NC_{ij}}{NC_i + NC_j} \tag{4}$$

NC_i, NC_j は作業者 T_i, T_j が付与した非コンテンツ領域の数，$NC_{i,j}$ は 2 人の作業者がともに付与した非コンテンツ領域の数である．一致度は，非コンテンツ領域単位とテキストユニット単位の両方で評価した．後者の場合は B または I のラベルを区別せず，非コンテンツ領域と判定した TU がどれだけ一致しているかを評価した．表 1 に示す作業者の一致度は十分高いとは言えない．これは非コンテンツ領域の判定が人によって揺れが生じやすいことを示唆する．非コンテンツ領域のより厳密な定義が必要であろう．これは今後の課題としたい．

3.2 結果

3.1 項で作成したデータを 5 分割し，1 つをテストデータ，残りを訓練データとする実験を 5 回繰り返す 5 分割交差検定を行った．データの分割は dmoz の登録ページとその子ページを 1 つの単位とした．したがってページ数は均等に 5 分割されていない．

実験結果を表 2 に示す．表 2 における LA は，提案手法の TU のラベル (B,I,O) の正解率を表わす．一方 LA_{bl} は，全ての TU のラベルを O としたベースラインシステムのラベルの正解率である．提案手法はベースラインシステムを大きく上回ることがわかる．

一方，R_{re}, P_{re}, R_{tu}, P_{tu} は非コンテンツ領域の検出に関する評価指標である．R_{re}, P_{re} はモデルが検出した非コンテンツ領域が正解データと完全に一致しているときを正解とみなしたときの再現率と精度 (適合率) である．一方 R_{tu} および P_{tu} は，TU を単位として評価した非コンテンツ領域の再現率と精度である．このとき，B または I ラベルが付与された TU はともに非コンテンツ領域とみなし，両者は区別しない．領域単位で評価した精度 P_{re} は 3 割程度と低いが，TU 単位で評価した精度 P_{tu} は約 7 割であった．これは，非コンテンツ領域を範囲を含めて完全に検出することは難しいが，部分的にはうまく検出できていることを示唆している．とはいえ，

表 2: 実験結果

LA	LA_{bl}	R_{re}	P_{re}	R_{tu}	P_{tu}	FP_{tu}
0.769	0.698	0.135	0.296	0.431	0.694	0.069

R_{tu} も P_{tu} も十分高いとは言えないため，手法の更なる改善が必要である．

FP_{tu} は，コンテンツ領域となるべき (O ラベルが正解となる)TU のうち，誤って非コンテンツ領域と判定された (モデルによって B または I ラベルが付与された) TU の割合である．本研究は様々なウェブアプリケーションの前処理と位置付けているため，コンテンツ領域を非コンテンツ領域と誤るのは有用な情報を切り捨てることになるために望ましくない．今回の実験では FP_{tu} は 7%と比較的低い値であることがわかった．

4 おわりに

本論文では，様々なウェブアプリケーションの前処理として，有用な情報を含まないウェブページの非コンテンツ領域を自動検出する手法を提案した．今後は，詳細なエラー分析を行い，非コンテンツ領域の検出に有効な新たな素性を発見し，再現率や精度を向上させたい．例えば，非コンテンツ領域の多くはページの上下左右に位置し，ページの中央に位置することは少ない．このようなレイアウト上の位置情報は有効な素性になりうる．また，現在は非コンテンツ領域か否かの識別しか行っていないが，非コンテンツ領域を「広告」「ナビゲーションリンク」「著作権表示」などのタイプに分類し，非コンテンツ領域を検出するとともにそのタイプも識別することも試みたい．

参考文献

[1] 加藤邦彦, 白井清昭. 視覚障害者用音声ブラウザのためのウェブページ解析. 言語処理学会第 12 回年次大会, pp. 809–812, 2006.

[2] Shian-Hua Lin and Jan-Ming Ho. Discovering informative content blocks from web. In *Proceedings of the the Eighth International Conference on Knowledge Discovery and Data Mining*, pp. 588–593, 2002.

[3] 中村達也. ウェブページにおける非コンテンツ領域の検出に関する研究. Master's thesis, 北陸先端科学技術大学院大学, 3 2007.

[4] 南野朋之, 齋藤豪, 奥村学. 繰り返し構造を用いた Web ページの構造化に関する研究. 自然言語処理研究会 2003-NL-154, pp. 185–192, 2003.

[5] Shipeng Yu, Deng Cai, Ji-Rong Wen, and Wei-Ying Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the the Twelfth International World Wide Web Conference*, 2003.