修 士 論 文

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

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

中村達也

2007年3月

修 士 論 文

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

指導教官　白井 清昭

審査委員主査　白井 清昭 助教授
審査委員　島津 明 教授
審査委員　鳥澤 健太郎 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

510073 中村達也

提出年月: 2007 年 2 月

概 要

近年の WWW の普及により, ウェブページから様々な情報を得る機会が多くなった. また, ウェブページを対象とした研究として, 情報検索やウェブマイニングなどが数多く行われている. しかし, ウェブページには情報として有用な領域 (コンテンツ領域) とそうでない領域 (非コンテンツ領域) が混在している. 情報として有用でない非コンテンツ領域は情報検索やウェブマイニングなどのアプリケーションに不要である. したがって, ウェブページ内の情報をコンテンツ領域と非コンテンツ領域とに識別できることが望ましい. そこで, 本稿では, 非コンテンツ領域が付与されたウェブページから素性ベクトルを抽出し, それらからチャンキングによって非コンテンツ領域を検出するモデルを学習することで, ウェブページ内の非コンテンツ領域を自動的に検出する手法を提案する. 実験により、提案手法によって約 7 割の精度で非コンテンツ領域を検出することができたことを確認した。

目 次

図 目 次

表 目 次

第1章　はじめに

近年WWWの普及により、ウェブページから様々な情報を得る機会が多くなってきた。これに伴い、情報検索、ウェブマイニング、ウェブコーパスの構築など、ウェブページを対象とした研究が数多く行われている。しかし、ウェブページには情報として有用なコンテンツ領域とそうでない領域 (非コンテンツ領域) が存在する。ここで非コンテンツ領域とは、情報発信を目的として作成されてない領域と定義する。例えば、広告、目次、検索フォームなどが該当する。このような非コンテンツ領域は、ウェブ検索、ウェブマイニングなどに悪影響を与えると考えられるため、自動的に検出することが望ましい。

非コンテンツ領域が情報検索やウェブマイニングに対して及ぼす悪影響の例をいくつか挙げる。情報検索を行う場合、ウェブページの単語に対して索引語付けを行い、ユーザからの検索クエリに含まれる単語とページの索引語集合を照合することで、ユーザが知りたい情報を含むページ検索をする。ここで、ウェブページに非コンテンツ領域が存在すると、非コンテンツ領域の単語も索引語付けをしてしまい不適切なページが検索されてしまうことが考えられる。例えば、図 1.1 のウェブページにおいて、実線で囲まれた部分はナビゲーション目的のリンクである。ある検索クエリから、この領域の『サイエンス』というキーワードがヒットしてこのページが検索されたとしても、このページが適切なページであるとは言えない。このように、非コンテンツ領域の単語を索引語として用いるべきではない。

また、ウェブマイニングに関しては、ウェブページの情報として有用でない非コンテンツ領域から知識獲得をすることは無駄である。また、ウェブページのテキストに対し形態素解析または構文解析などの自然言語処理を行ったりするとき、ウェブページ全体を解析するよりも、非コンテンツ領域を検出し、非コンテンツ領域を無視することで処理時間を短縮することができる。

このように、ウェブページの非コンテンツ領域を検出することで、様々なアプリケーションの性能の改善をすることができる。特に、情報検索に関してはウェブページの情報として有用でない非コンテンツ領域を検出することでその領域に現れる単語に対して索引語付けをしないようにすると、不適切なページが検索されることが無く、情報検索の精度を向上させることができると考えられる。

これらのことから、本研究ではウェブページにおける情報として有用でない非コンテンツ領域を検出するシステムを提案する。ページのどのような部分を非コンテンツ領域とするかはアプリケーションによって異なるが、本研究では情報検索を想定する。提案手法では、ウェブページの HTML タグで分割されたテキストに対して非コンテンツ領域を構成

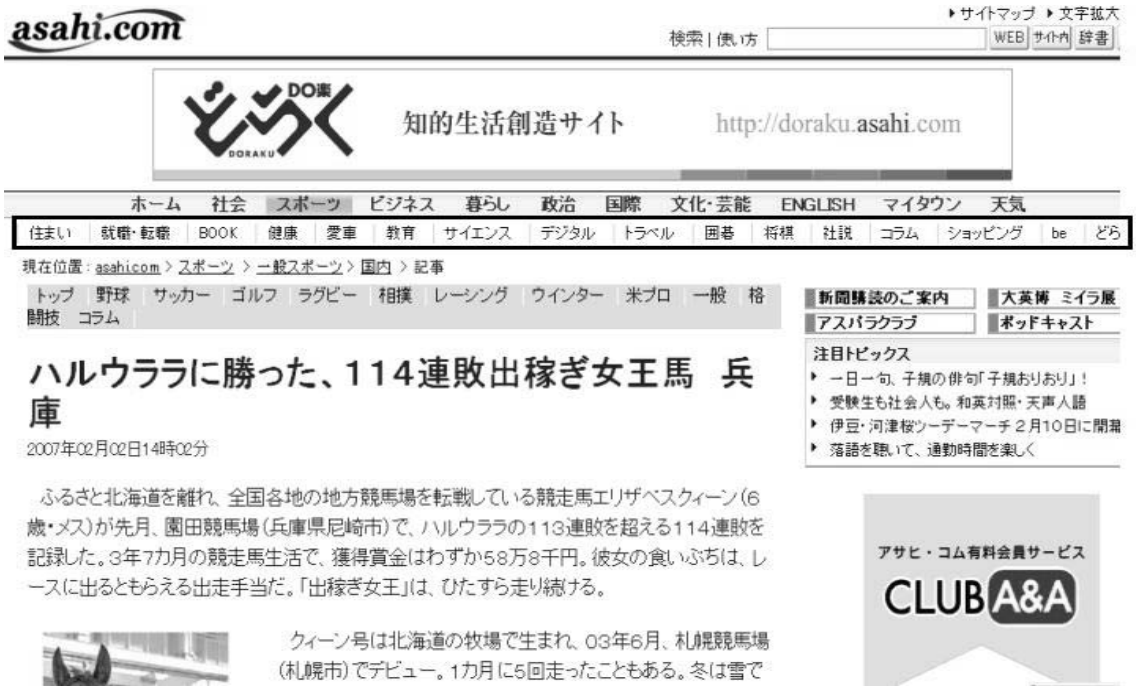

図 1.1: 非コンテンツ領域を含むウェブページ

するテキストか、コンテンツ領域を構成するテキストかのラベル付けを行う。そのラベル付けは、正解として非コンテンツ領域の付与されたウェブページから学習したモデルを用いて行う。学習には、汎用チャンキングツール YamCha[6] を用いる。そして、非コンテンツ領域を構成するテキストとラベル付けされたテキストを 1 つにまとめあげることで非コンテンツ領域を検出する。

本論文の構成は、以下の通りである。第 2 章では、本研究の関連研究について述べる。第 3 章では、本研究で提案するウェブページにおける非コンテンツ領域の検出手法について述べる。第 4 章では、提案システムの有効性を確認するために行った実験とその結果について述べる。第 5 章では、本研究についてのまとめと今後の課題について述べる。

第2章　関連研究

本章では、本研究の関連研究について述べる。2.1 節では、ウェブページのコンテンツ検出に関する研究について述べる。2.2 節では、ウェブページの構造解析に関する研究について述べる。

2.1　ウェブページのコンテンツ検出に関する研究

Lin らは、まず同ウェブサイトのテンプレートが同じであるウェブページを HTML タグ<table>に従っていくつかのコンテンツブロックに分割し、そのブロックのエントロピーをそこに現れる素性 (単語) のエントロピーから算出し、分割されたブロックが情報として有用かどうかを判定する InfoDiscoverer というシステムを提案している [1]。分割されたブロックがウェブページの情報として有用かどうかの判定は動的に選択されたエントロピーの閾値を用いている。この提案システムにより、サイトの情報として有用なブロック (ニュース記事) と意味的に冗長なコンテンツ (広告、バナー、ナビゲーションパネルなど) を自動的に分けることができる。また、この処理を情報検索や抽出アプリケーションの前処理として行うことで、情報検索の精度向上やインデックス付けのサイズの減少、抽出の複雑さを減らすことができると述べている。ただし、この研究で扱うウェブページは同一サイトのテンプレートが同じページを対象にしている。これに対し、本研究では、様々なウェブページに対して情報として有用なコンテンツ領域とそうではない非コンテンツ領域の検出を目的としている。

Debnath らは、ウェブページには、情報として有益でないコンテンツとしてナビゲーションリンク、広告、コピーライト表示、ロゴ、カウンタ、検索フォームなどがあると述べている。ここで、知的情報処理システムの検索エンジンがウェブページの索引語付けを行うとき、これらのコンテンツが存在することで、情報として有益なコンテンツのみに対して索引語付けができないことを問題としている [4]。そのため、ウェブページの情報として有用なコンテンツブロックと情報として有用でないブロックを分けることがとても重要であると述べている。この二つのブロックを同定するためのアルゴリズムとして提案されているのが FeatureExtractor である。このアルゴリズムは、HTML タグ<table>や<tr>,<p>,<hr>,<ul>のようなタグに従ってブロックに分けている。そしてブロックのテキストの素性に基づいて有用かそうでないかを同定する。この研究は、ルールベースで非コンテンツ領域を検出するアルゴリズムである。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考えられるが、アプリケーション毎に別のルールセットを用意しな

ければならない可能性が高く、様々なアプリケーションに対応しずらい。これに対し、本研究では、正しい非コンテンツ領域が付与されたウェブページを正解データとして用意し、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。そのため、正解データを変えることで様々なアプリケーションに対して柔軟に対応できる。

2.2 ウェブページの構造解析に関する研究

Yu らは、ウェブページは一般に複数のトピックを含んでいるため、ページ全体で擬似関連フィードバックを行うと、意味的に関連のないキーワードに高い重みを付けてしまうことを問題としている [2]。そこでページを DOM の情報と視覚情報を組み合わせて用いることでページの意味構造を解析し、ページをトピック毎に領域に分割する VIPS アルゴリズムを提案している。キーワードを含む領域のみに擬似関連フィードバックを用いることで、広告やリンクなどの無関係な領域の単語は無視し、意味的に関連のあるキーワードのみに重みをつけ、情報検索の性能を改善した。これは、ページにおける情報として有用でない非コンテンツ領域を検出することが情報検索に役立つことを意味している。

Yang らは、ウェブページに関してコンテンツの視覚的な類似性を検出することでページの意味構造を自動的に解析するアプローチを提案している [5]。ほとんどのウェブページに関して、同じ内容のカテゴリーにはサブタイトルやレイアウトスタイルに一貫性があり、異なったカテゴリー間には明白な境界線がある。ウェブページにおいて、視覚的な類似性がある頻繁に現れるパターンを検出するアルゴリズムを適用し、それらから最も適切なパターンを選ぶために多くのヒューリスティックを用いる。それらのパターンに従い項目をグループ化することによって、ウェブページの階層的な表現 (tree) として視覚的な一貫性を構築できる。

加藤らは、視覚障害者がウェブページの情報を得る場合、ページの文字情報を音声で読み上げるツールを用いているが、それでは知りたい情報が書かれた箇所をツールで読み上げるまでに手間と時間がかかるという問題があるとしている。そこで、不要な情報の読み飛ばしをするシステムを開発するために、ウェブページの構造解析を行い、ページのセグメンテーションを行う手法を提案している [8]。このセグメンテーション手法において、検出すべきセグメントを次のように定義している。(1) 意味のあるまとまり (共通の属性をもつ) 、(2) レイアウト上のまとまり、(3) 要素が 2 つ以上ある、(4) セグメントの先頭に見出しがあることが望ましい、(5) 各セグメントは入れ子になってよい、(6) ページ全体をカバーしなくてもよい。このように定義されたセグメントを検出するために、ウェブページの DOM 構造を用いて大まかなセグメンテーションを行う。具体的にセグメンテーションに用いられるタグとして、タグで囲まれている箇所をセグメントとみなす table,ol,dl,ul,p や、タグ自体をセグメントの境界とみなす h1,h2,h3,h4,h5,h6,hr を使用している。そして、DOM 構造を用いたセグメンテーションがしばしば大きすぎるため、そのセグメントの分割を行う。その分割手法として、イメージによる分割とテーブル内の部分木を利用した分割を行う。これによって検出されたセグメントは最小単位のセグメントとなる。また、リ

ストタグのようなセグメントはヘッダ部分を含まないことがある。しかし、そのヘッダ部分はセグメントの内容を表すことがあり、読み飛ばしを行うときにはヘッダ部分を読み上げる。従って、そのヘッダ部分とセグメントを一つにマージすることで新たなセグメントを作成している。

南野らは、ウェブ上の情報は、レイアウト記述言語で記述されて、人が目で理解するための情報であるため、それを計算機で直接扱うのは困難な点があるとしている。そこで、人間がページのレイアウトから理解する構造に近い形でウェブ上の情報を計算機が扱うための手法を提案している [9]。彼らの手法は、HTML 文章中に含まれるタグの繰り返し構造から、HTML 文章の構造認識を行い自動的なセグメンテーション、構造化をすることを目的としている。まず、最下層のタグの繰り返しから構造化をする。例えば、ページの目次などは<a>タグが張られたテキストである。よって、目次には<a>タグの繰り返しが存在し、その部分が構造として検出されることになる。そして、その構造をグループ化し、さらに上の階層のタグの繰り返し構造の検出を行う。このように、ボトムアップに検出を繰り返すことで HTML 文章の構造化を行う。しかし、繰り返しによる構造化がページ全体に対して行えないことがある。そのときは、DP マッチングを用いてタグの類似性からページ全体に対しての構造化を行う。このように、ウェブページの自動的なセグメンテーション、構造化ができれば、ウェブ上の情報を扱う様々なアプリケーションにとって有用である。

本研究は、ウェブページのテキストが非コンテンツ領域かどうか判定して、そのテキストをまとめ上げることで非コンテンツ領域の検出を行っている。したがって本研究は、上に挙げたウェブページの構造解析に関する研究 [2, 5, 8, 9] とはその目的が異なるが、ウェブページの意味的なまとまりを検出する点で関連がある。ただし、本研究ではウェブページの構造化は行っていない。しかし、DOM ツリーを利用するなど、ウェブページ解析に用いる手がかりがこれらの研究と共通するものがある。

第3章　非コンテンツ領域の検出

この章では、ウェブページにおけるコンテンツ領域と非コンテンツ領域の定義を述べ、非コンテンツ領域を自動的に検出する手法とそれに用いる素性を示す。

3.1　非コンテンツ領域とは

非コンテンツ領域とコンテンツ領域の一般的な定義を以下に示す。

- 非コンテンツ領域の定義 ··· ウェブページ作成者が情報発信を意図していない領域
- コンテンツ領域の定義 ··· ウェブページの情報として有用な領域

すなわち、ウェブページにおける非コンテンツ領域とは、特に有用な情報を含まない領域というのが一般的な定義である。しかし、非コンテンツ領域の厳密な定義は、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、ナビゲーション目的のリンクグループなどは、情報検索の立場から考えると、誤ったページを検索してしまう原因となってしまうので必要ない。しかし、ウェブにおけるリンク構造を解析するときは、ナビゲーション目的のリンクは重要である。

このように、非コンテンツ領域の定義はアプリケーション毎に異なるが、本論文では、アプリケーションとして情報検索を想定し、非コンテンツ領域を定義する。すなわち、情報検索にとって有用かどうかという観点で非コンテンツ領域を定義する。具体的には、以下のいずれかに該当する領域を非コンテンツ領域と定義する。

- 非コンテンツ領域の定義
  - **検索フォーム**
    図 3.1 の B は検索フォームである。このような検索フォームに含まれる単語にヒットして検索されたページは適切なページである可能性が低いため、非コンテンツ領域とする。
  - **広告**
    図 3.1 の A と C、図 3.2 の F は広告である。このような広告はページの内容と異なっているので、この広告に含まれる単語でページが検索されてもそのページが適切であると言えない。したがって、非コンテンツ領域とする。

– **コピーライト表示**
図 3.2 の I はコピーライト表示である。コピーライト表示はページの内容とは無関係であり、この領域に含まれる単語にヒットして検索されたページが適切なページである可能性は低いため、非コンテンツ領域とする。

– **目次**
目次とは、ここではページの上部などにそのページの内容を目次のように箇条書きで紹介している領域を指す。また、目次はページ内リンクで構成されていることも多い。目次の中のキーワードは、それと同じものがページ内の他の箇所に存在すると考えられる。したがって、非コンテンツ領域として検出しその領域内の単語を索引語としなくても、情報検索に悪影響を及ぼすことはないと考えられる。したがって、目次は非コンテンツ領域とする。

– **ナビゲーション目的のリンクグループ**
図 3.2 の E と G と H はリンクグループである。これは、ナビゲーション目的のリンクが集まったものである。ある人が情報検索をしたとき、リンクグループに含まれる単語によって、このページが検索されるよりも、リンク先のページが検索されることが望ましいことが多いと考えられる。したがって、ナビゲーション目的のリンクグループは非コンテンツ領域とする。ただし、リンクの集合で構成されたリンク集は、このような非コンテンツ領域のリンクグループとして考えていない。リンク集とは、ページ作成者が意図的に関連のあるリンクを一つのページにまとめられたものである。よって、リンク集はページの情報として有用なものであると考え、非コンテンツ領域としなかった。

– **カウンタ**
カウンタはページを閲覧した人数を表示したものであり、検索するためのキーワードとして適切ではないことは明らかである。したがって、非コンテンツ領域とする。

– **サイトマップ**
サイトマップはサイト内のページへのリンクをまとめた領域である。ナビゲーション目的のリンクグループと同様に、サイトマップに含まれる単語でサイトマップのページが検索されるよりは、リンク先のページが検索される方が適切であると考えられるので、サイトマップは非コンテンツ領域とする。

3.2 検出手法

3.2.1 チャンキング

本研究では、図 3.3 のように矢印で指し示された HTML タグで挟まれたテキストに対してコンテンツ領域か非コンテンツ領域かのラベル付けを行う。

図 3.1: 非コンテンツ領域 (広告、検索フォーム) の実例

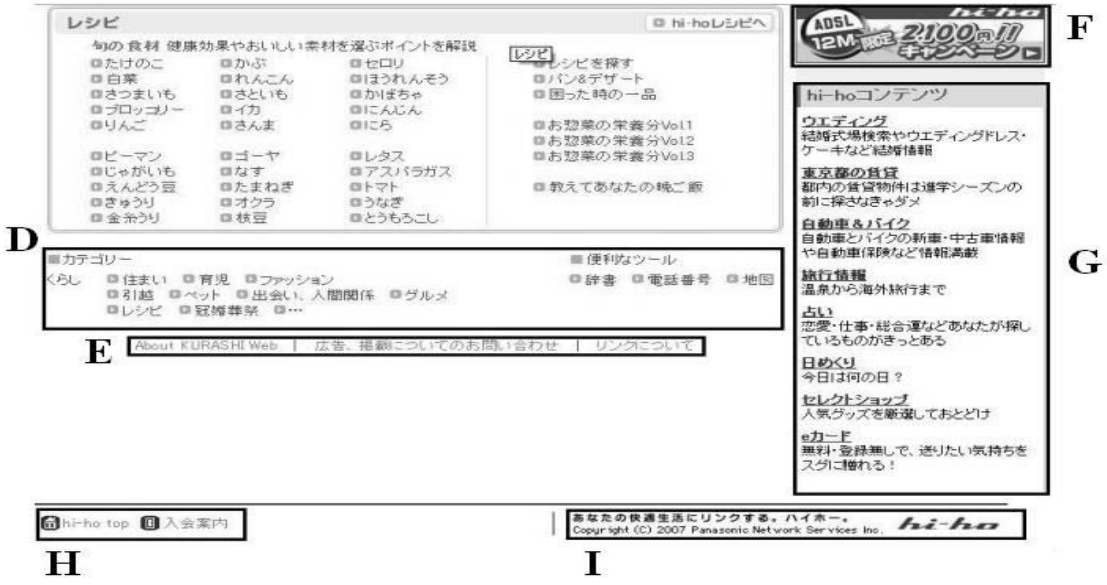

図 3.2: 非コンテンツ領域 (広告、コピーライト表示、リンクグループ) の実例

図 3.3: HTML タグで分割されたテキスト

ウェブページを図 3.3 のようなテキストに分割したとき、、非コンテンツ領域は複数のテキストから構成されていることから、チャンキングによって非コンテンツ領域内のテキストであるとラベル付けされたテキストをまとめ上げることで非コンテンツ領域を検出する手法を提案する。コンテンツ領域か非コンテンツ領域かのラベル付けを行うチャンキングのモデルは、非コンテンツ領域があらかじめ付与された正解付きのウェブページから学習する。本手法では、チャンカーとして YamCha を使用した。YamCha は機械学習アルゴリズムとして SVM(Support Vector Machine) を用いてチャンキングのモデルの学習を行っている [6]。SVM は、他の学習モデルと比較すると極めて汎化能力が高く、高次元の素性集合を用いても過学習しにくいという特性を持っている。

3.2.2 ラベル

本手法では、図 3.3 のように HTML タグで分割されたテキストに対してコンテンツ領域、非コンテンツ領域のチャンキングのラベル付けを行う。そのラベルのモデルとして図 3.4 のようにテキストにラベル付けする IOB2 モデルを用いている。IOB2 モデルとは、チャンクである非コンテンツ領域の先頭のテキストに対してラベル B を付ける。また、非コンテンツ領域内の先頭に位置するテキスト以外のテキストに対してはラベル I を付ける。そして、非コンテンツ領域以外すなわちコンテンツ領域のテキストに対してラベル O を付ける。

- IOB2 モデル

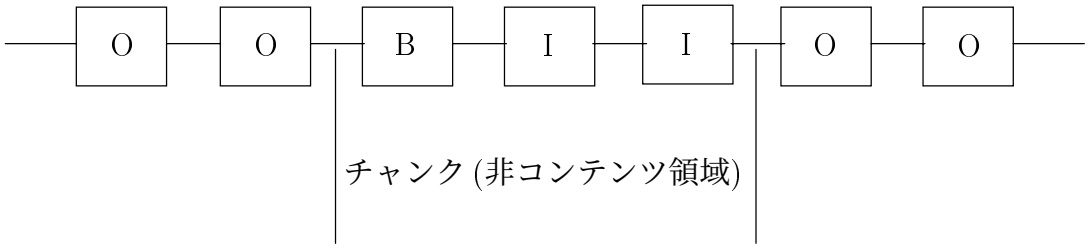

図 3.4: IOB2 モデルによるラベル付けの例

B ··· チャンク (非コンテンツ領域) の先頭を示すテキスト I ··· チャンク内のテキスト (先頭以外) O ··· チャンク以外のテキスト

正解付きデータから、ウェブページ内の各テキストのラベルが B、I、O のいずれであるかを判定するモデルを学習する。未知のウェブページに対してテキストのラベルの判定を行い、B を先頭とし I が続くテキスト集合が非コンテンツ領域として検出される。

3.3 素性

コンテンツ領域か非コンテンツ領域かを区別するための素性を以下に挙げる。これらの素性は、事前に採集した 21 ページのウェブページを対象に、非コンテンツ領域の検出に有効な手がかりを調査し、設定した。チャンキングのモデルを学習する際には、正解データ内のテキストに対して素性ベクトルを作成し、学習データとした。

- 非コンテンツ領域に現れやすいキーワード
  テキストに含まれるキーワードに着目し、非コンテンツ領域に出現しやすいキーワード (名詞) がテキストに含まれているかどうかを素性として、非コンテンツ領域かコンテンツ領域かの判定をする。非コンテンツ領域に出現しやすいキーワードの例としては『ホーム』、『Copyright』、『TOP』などが挙げられる。非コンテンツ領域に含まれているキーワード『ホーム』、『Copyright』の例を図 3.5 に示す。このようなキーワードは、様々なウェブページの非コンテンツ領域内のテキストに現れることが多い。このことからテキストに『ホーム』、『Copyright』、『TOP』といった非コンテンツ領域によく出現するキーワードが含まれるかどうかを素性として用いる。

  本研究では非コンテンツ領域内によく出現するキーワードを自動的に選別している。その選別条件を以下に示す。

図 3.5: 非コンテンツ領域に含まれるキーワード

– 非コンテンツ領域内のテキストによく出現するキーワードの選別条件
まず、学習データにおける、キーワード w が非コンテンツ領域に出現する確率 $P_{nc}(w)$ を式 (3.1) から計算することで非コンテンツ領域によく出現するキーワードを素性として選別する。そのとき、学習データに現れる合計回数が 20 回以上であり、かつ $P_{nc}(w)$ が 0.7 以上であるキーワード w を素性として選別する。しかし、学習データでの出現回数と $P_{nc}(w)$ が高いキーワードであっても、そのキーワードが一般的に非コンテンツ領域を示唆するようなものでないことがある。例えば、あるドメインのウェブページの非コンテンツ領域にのみよく出現するキーワードは、例え $P_{nc}(w)$ が高くても非コンテンツ領域に頻出する一般的なキーワードとは言えない。

そこで、ウェブページの URL のドメインに着目し、キーワードが含まれるウェブページのドメインの異なり数 D_w をキーワード w が非コンテンツ領域に現れる確率 $P_{nc}(w)$ に掛ける (式 (3.2))。ここで D_w が大きければ、キーワード w が様々なドメインのウェブページにおける非コンテンツ領域に出現することを意味する。

$$P_{nc}(w) = \frac{\text{キーワード } w \text{ の非コンテンツ領域における出現回数}}{\text{学習データにおけるキーワード } w \text{ の出現回数}} \tag{3.1}$$

$$P_{nc}(w) \times D_w \tag{3.2}$$

このように、ページの非コンテンツ領域に多く出現するキーワードの確率が高いだけでなく、様々なドメインのウェブページに現れるキーワード、すなわちどのサイトのページにおいても非コンテンツ領域内に含まれるようなキーワードに対して式 (3.2) の値が大きくなる。式 (3.2 の値の大きい単語を選別することで素性として相応しいキーワードを特定できる。以上のことからキーワードの選別条件をまとめると、

* キーワード w が学習データに現れる合計回数が 20 回以上
* キーワード w が学習データの非コンテンツ領域に現れる確率 $P_{nc}(w)$ が 0.7 以上
* $P_{nc}(w)$ に D_w を掛けた結果が 2 以上

となる。この条件を満たすキーワードを含むか否かを素性とする。

- テキスト長
  図 3.6 の実線で囲まれた非コンテンツ領域に含まれるテキスト『お問合せ』と、破線で囲まれたコンテンツ領域に含まれるテキスト『複数ドメインをお持ちの方、サーバー費用を節約できます』を比較すると、ページの本文となる領域のテキスト長は長く、ページの目次となっている部分のテキスト長は短いことが分かる。本研究において、テキスト長とはテキストを構成する文字数である。このように、テキスト長は非コンテンツ領域の検出に有効な素性である。ただし、テキスト長そのものの値を素性としてしまうとデータの過疎性の問題が生じやすい。そのため素性の値を表 3.1 のように与えた。

表 3.1: テキスト長の素性値の与え方

$x = 1 \rightarrow$ `one`
$x = 2 \rightarrow$ `two`
$3 \leq x \leq 5 \rightarrow$ `three_five`
$6 \leq x \leq 8 \rightarrow$ `six_eight`
$9 \leq x \leq 15 \rightarrow$ `nine_fifteen`
$16 \leq x \rightarrow$ `over_sixteen`
ここで x はテキスト長である

この素性により、テキスト長が長いものがコンテンツ領域を、テキスト長が短いものは非コンテンツ領域を構成しているという傾向を学習することが可能となり、非コンテンツ領域の判定に有効であると考える。

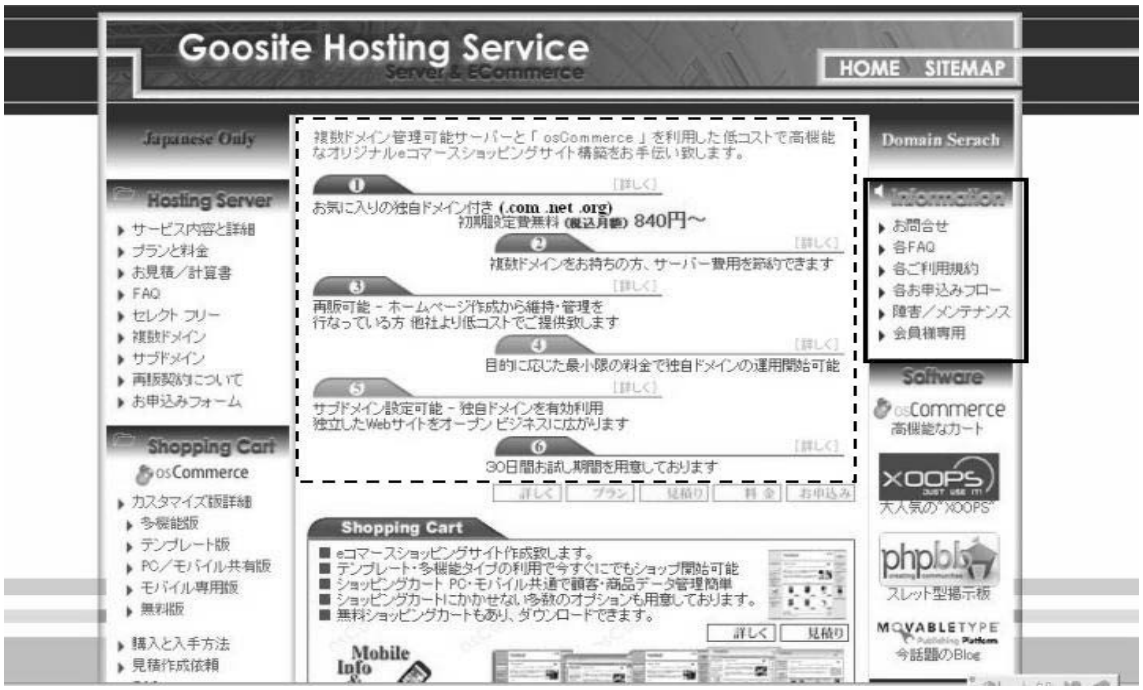

図 3.6: テキスト長

- 動詞、形容詞があるか
  図 3.7 において、実線で囲まれている非コンテンツ領域に含まれるテキスト『ホーム』や『住まい』には動詞、形容詞が含まれていない。一方、破線で囲まれたコンテンツ領域に含まれるテキスト『３人が死亡、５人が負傷した･･･業務上過失致死傷容疑で逮捕した。』には『怠った』や『する』といった動詞が含まれている。このように、ページの本文となるようなテキストは動詞や形容詞 (さらには文) を含むことが考えられる。逆に目次などページの情報として有用でない非コンテンツ領域内のテキストは動詞や形容詞を含んでおらず、テキストが文になっていないことが多い。このことから、テキストに動詞、形容詞が含まれているかいないかを素性として加えて領域の判定に用いる。

- DOM ツリーの末端が<a>タグか (内部リンクか外部リンクか？)
  図 3.8 において、ウェブページの本文となる点線で囲まれたコンテンツ領域のテキスト『「納豆で減量」実験捏造･･･フジ系「あるある大事典」(以下省略)』にはリンクが張られておらず、実線で囲まれた非コンテンツ領域のテキスト『女優気分泡に包まれ･･･』や『疲れた体をじわっと･･･』などにはリンクが張られている。この例のように、非コンテンツ領域におけるナビゲーション目的のテキストにはリンクが張られていることが多い。そこで、テキストにリンクが張られているかどうかを素性として与える。さらに、リンクが張られている場合は内部リンクか外部リンクかを判断し素性とする。ここで、内部リンクとは同一サイトへのリンクであり、外部

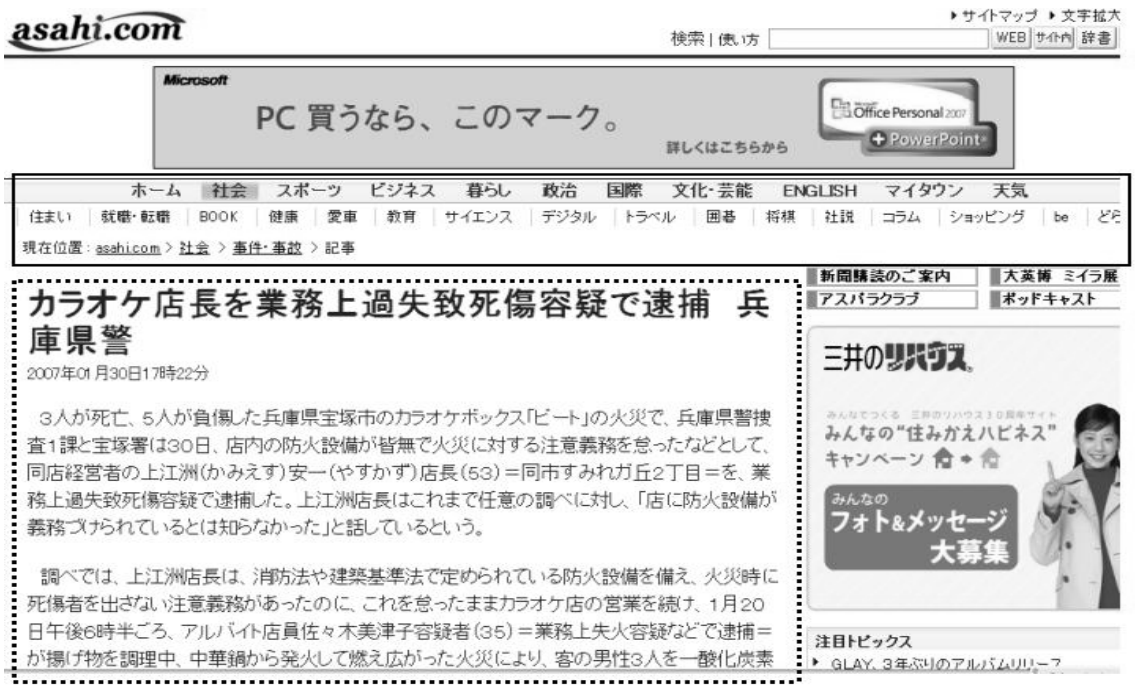

図 3.7: 動詞、形容詞があるか

リンクはそれ以外のサイトへのリンクと定義する。内部リンクか外部リンクかの判定は次のように行う。

- ページの URL のドメインとリンクの URL のドメインが同じ
  または、リンクの URL が相対パス → **内部リンク**

- ページの URL のドメインとリンクの URL のドメインが異なる → **外部リンク**

ナビゲーション目的のリンクは、非コンテンツ領域であるが、そのほとんどは内部リンクであると考えられる。したがって、内部リンクのアンカー内にあるテキストは非コンテンツ領域である可能性が高い。このようにテキストにリンクが張られているかどうか、張られているリンクが内部リンクか外部リンクかは非コンテンツ領域の判定のための素性として有効であると考えられる。

- DOM ツリー上で近傍にある HTML タグ
  テキストは DOM ツリーの葉の部分に位置していて、そのテキストがどのようなタグの下にあるのかという情報を素性とする。具体的には、テキストからルートに DOM ツリーを辿ったとき、テキストに近い位置にある 3 つの HTML のタグを素性として用いている。しかし、文章装飾タグなどの HTML の構造に関係のないタグは素性として用いていない。このようなタグは、HTML の構造を表すものではないのでテキ

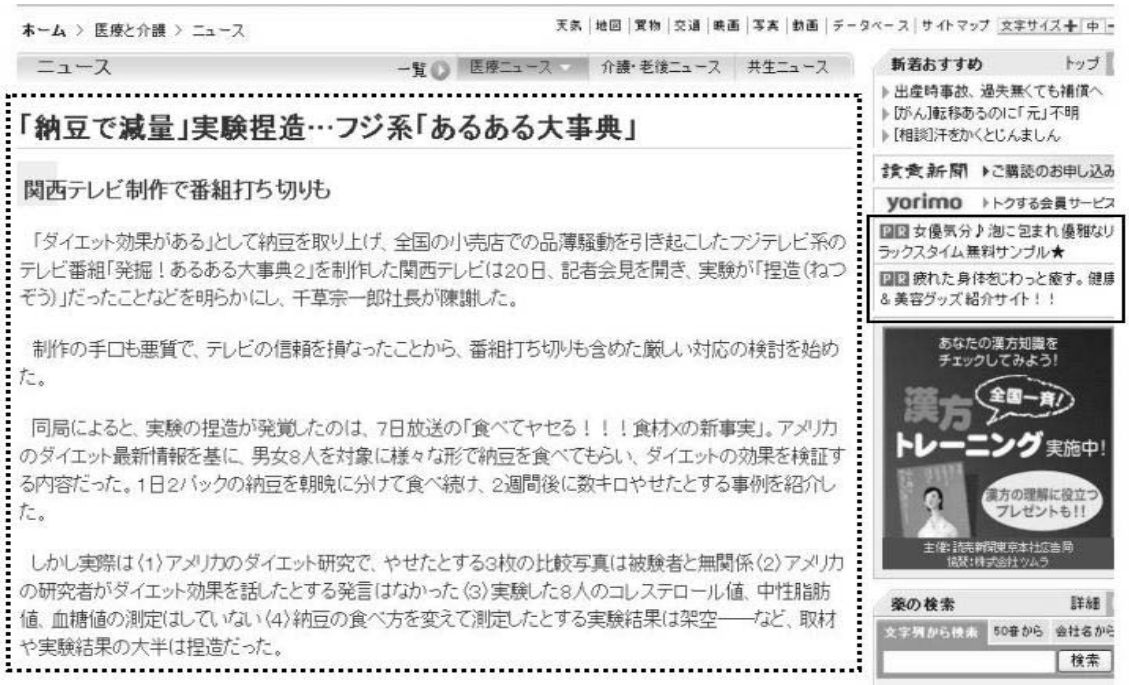

図 3.8: リンクの有無

ストの位置を示す素性として有効ではない。今回、HTML の構造に関係のないタグとして素性として加えなかったタグは、div、font、a、span、strong、select、option、pre、small、kbd である。これらのタグの効果の例を表 3.2 に示す。

目次やリンクグループは<table>タグやリストタグ (<ol>,<ul>,<dl>,<li>) から構成されていることが多い。例えば、図 3.9 の場合、実線で囲まれた非コンテンツ領域内のテキスト『ホーム』はリストタグ<ul>内にあり、『ホーム』の上 3 つのタグは図 3.10 のように<body>,<ul>,<li>となる。また、破線で囲まれたコンテンツ領域内のテキスト『世界に羽ばたく科学者、技術者を育てる』や『本学は、科学と技術の分野で世界最高水準の研究と教育を … (以下省略)』は<table>やリストタグ内にはなく、テキストの上 3 つのタグは図 3.10 から<html>,<body>,<p>となる。このように、テキストが DOM ツリーでどこの位置にあるかの情報は非コンテンツ領域の判定に有効な素性であると考えられる。

- 直前のテキストに比べて DOM のパスが深くなるか浅くなるか
  図 3.11 の実線で囲まれたナビゲーション目的の非コンテンツ領域と破線で囲まれた本文のコンテンツ領域を見ると視覚的にも意味的にもこの領域の間に境界があるのがわかる。そこで、テキスト『最新情報』と直前のテキスト『共済・会員サービス』を見ると、DOM ツリーにおけるルートからテキストまでのパスは、『最新情報』が html-head-body-ul-li となり、『共済・会員サービス』は html-head-body-h2 となって

表 3.2: HTML の構造と関係のないタグ

`div`：テキストの特定の範囲について色などの変更を設定
`font`：テキストの色、サイズなどを指定
`a`：リンクの設定
`strong`：テキストを強調
`select`：プルダウン形式のメニューの作成
`option`：プルダウン形式のメニューの選択肢
`pre`：入力したとおりのテキストを出力
`small`：テキストのフォントを小さめに指定
`kbd`：ソースコードや出力結果を表示

図 3.9: テキストの近傍の HTML タグ

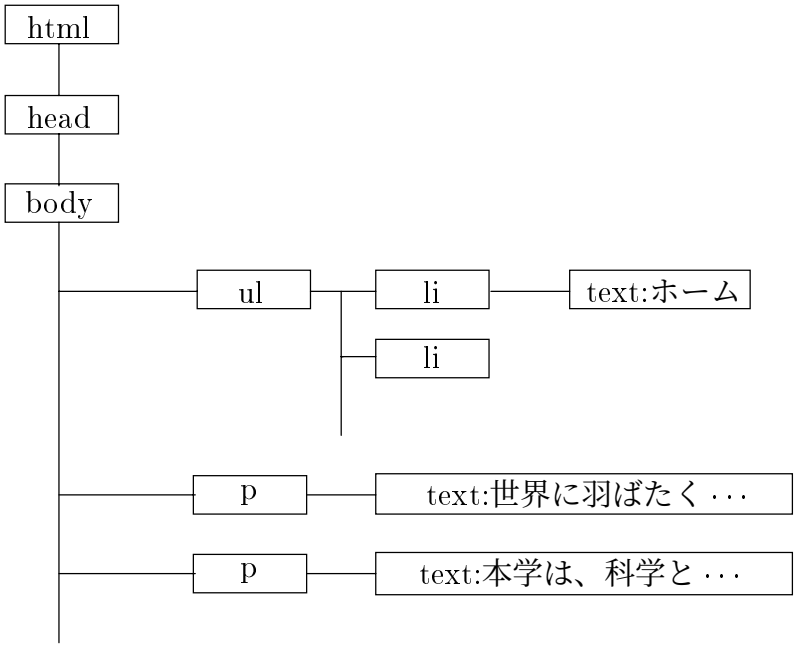

図 3.10: 図 3.9 の DOM ツリー

いる。このように、この領域境界部分で、直前のテキストと比較してルートからテキストまでのパスの深さに変化が起こっていることがわかる。これを素性に加えることによって、領域の境界を認識することができるのではないかと考えた。素性がとる値としては表 3.3 のように与えた。また、ページの先頭のテキストの素性値は shallow と与えた。

表 3.3: DOM のパスの変化の素性値の与え方

直前のテキストの DOM のパスと同じ → same
直前のテキストの DOM のパスと比べ浅くなる → shallow
直前のテキストの DOM のパスと比べ深くなる → deep

図 3.11: DOM のパスの深さの変化

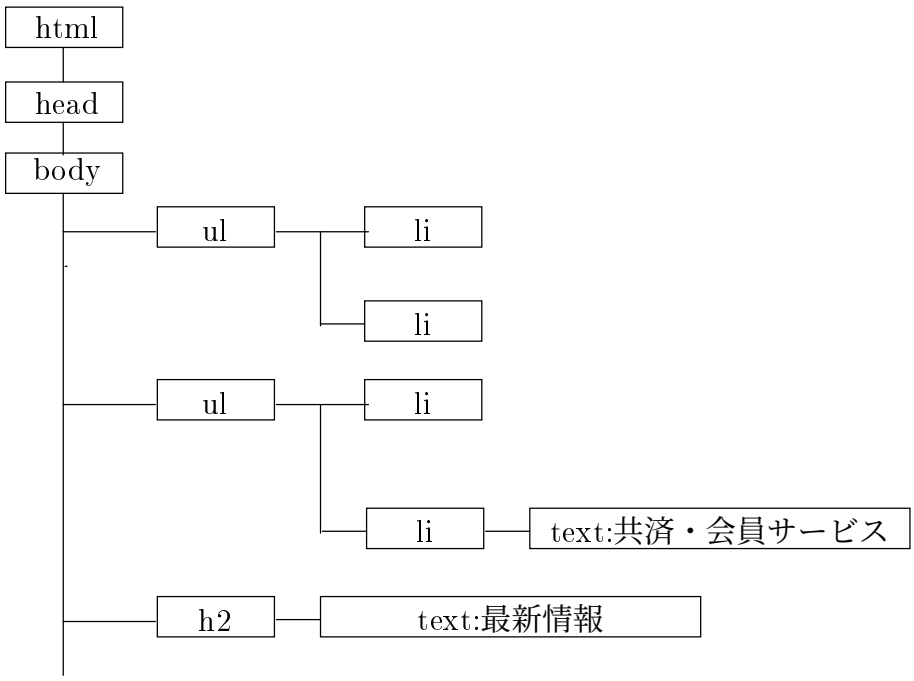

図 3.12: 図 3.11 の DOM ツリー

- <table>タグ内のテキストの平均長
  <table>タグ内に含まれるテキスト長の平均を素性とする。図 3.13 の実線で囲われた領域は、ページの目次であり、非コンテンツ領域である。一方、破線で囲われた領域はページの本文であり、コンテンツ領域である。ここで、図 3.14 は図 3.13 の DOM ツリーで、実線、破線で囲まれた部分は図 3.13 と対応する。この図 3.14 の実線領域と破線領域の<table>の下に位置するテキストの長さの平均を求める。実線で囲まれている目次などの非コンテンツ領域では、テキスト長が短いので平均長も短くなり、破線で囲まれているコンテンツ領域のテキスト長は長く平均長も長くなる。このようにして、目次などの非コンテンツ領域と本文のコンテンツ領域を区別することができる。

  一方、<table>の下のテキスト長が短いものと長いものが混在する場合がある。その例を図 3.15 に示す。この図の実線で囲まれた部分は、<table>タグで構成されている。この中のテキストは、ナビゲーション目的のリンクの非コンテンツ領域である。この領域のテキストは、テキスト長が短いものが多い。しかし、『オンラインショッピングに関するお問い合わせ』のテキスト長は 21 と長い。このために、素性としてテキスト長だけを用いてラベルの判定をすると、コンテンツ領域と判定されてしまうことが考えられる。そこで、<table>タグ内のテキストの平均長を素性として用いる。このとき、この<table>タグの下にあるテキスト全てに同じ素性値、しかも小さい値が与えられるため、このような例外的な領域のテキストに対しても正しく非コンテンツ領域と判定されることが期待できる。

  <table>タグ内のテキストの平均長を素性値としてそのままの値を与えるとデータの過疎性の問題が生じることが考えられる。よって、素性の値は表 3.4 のように与えた。

表 3.4: テキスト平均長の素性値の与え方

$x = 0 \rightarrow$ `zero`
$x = 1 \rightarrow$ `one`
$1 < x < 4 \rightarrow$ `one_four`
$4 \leq x \rightarrow$ `over_four`
ここで x はテキストの平均長である

- <table>タグ内のリンクの割合
  <table>タグ内のリンクの割合とは、<table>タグ内に含まれるアンカーテキストの割合のことを意味している。再び図 3.13 と図 3.14 における、実線部分 (非コンテンツ領域) と破線部分 (コンテンツ領域) に注目する。ここで、図 3.14 の非コンテンツ領域の実線で囲まれた部分のテキストには<a>タグによってリンクが貼られていることがわかる。一方、コンテンツ領域の破線で囲まれた部分のテキストには<a>タグが

なくリンクが貼られていないことがわかる。このように、<a>タグに着目し<table>タグ内のアンカーテキストの割合を見ると、非コンテンツ領域では 8 個のテキスト中アンカーテキストは 8 個であるので割合は 1 となり、コンテンツ領域での割合は 3 個のテキストのうちアンカーテキストは 0 個であるので割合は 0 となる。

このようなリンクの割合は、非コンテンツ領域の検出に有効な素性と考えられる。さらに、図 3.16 の実線で囲まれているナビゲーション目的のリンクに注目する。この領域内のテキストで、『INDEX』だけがアンカーテキストではない。このため、『INDEX』はコンテンツ領域のテキストであると判定される可能性がある。ところが、同じ<table>内に含まれる多くのテキストがアンカーテキストであることから、『INDEX』も非コンテンツ領域と判定できる。<table>タグ内のリンクの割合という素性を導入することにより<table>全体をまとめて非コンテンツ領域と判定する効果が期待できる。すなわち、リンクの割合が高いということで『INDEX』が非コンテンツ領域に含まれるテキストであるという判定ができるのではないかと思われる。

<table>タグ内のリンクの割合の値をそのまま用いてしまうとデータの過疎性の問題が生じる。そこで、素性の値を表 3.5 のように与えた。

表 3.5: リンクの割合の素性値の与え方

$x = 0 \rightarrow$ zero
$0 < x < 0.4 \rightarrow$ zero_zero-dot-four
$0.4 \leq x < 0.6 \rightarrow$ zero-dot-four_zero-dot-six
$0.6 \leq x < 1 \rightarrow$ zero-dot-six_one
$x = 1 \rightarrow$ one
ここで x はリンクの割合である

このようにして、<table>タグ内のリンクの割合を見ることで、コンテンツ領域と非コンテンツ領域の区別をつけることができる。

- 直前のテキストに付与されているラベル (B、I、O のいずれか)
  非コンテンツ領域は、複数のテキストから構成されている。このことから、直前のテキストに B、I、O のどのラベルが付与されているのかを見ることで、その次のテキストにどのラベルが付与されていることが多いかという情報が得られる。この情報は、チャンキングの素性として有用なものと考えることが出来るので、直前のテキストに付与されたラベルを素性とした。

図 3.13: テキスト平均長とリンクの割合

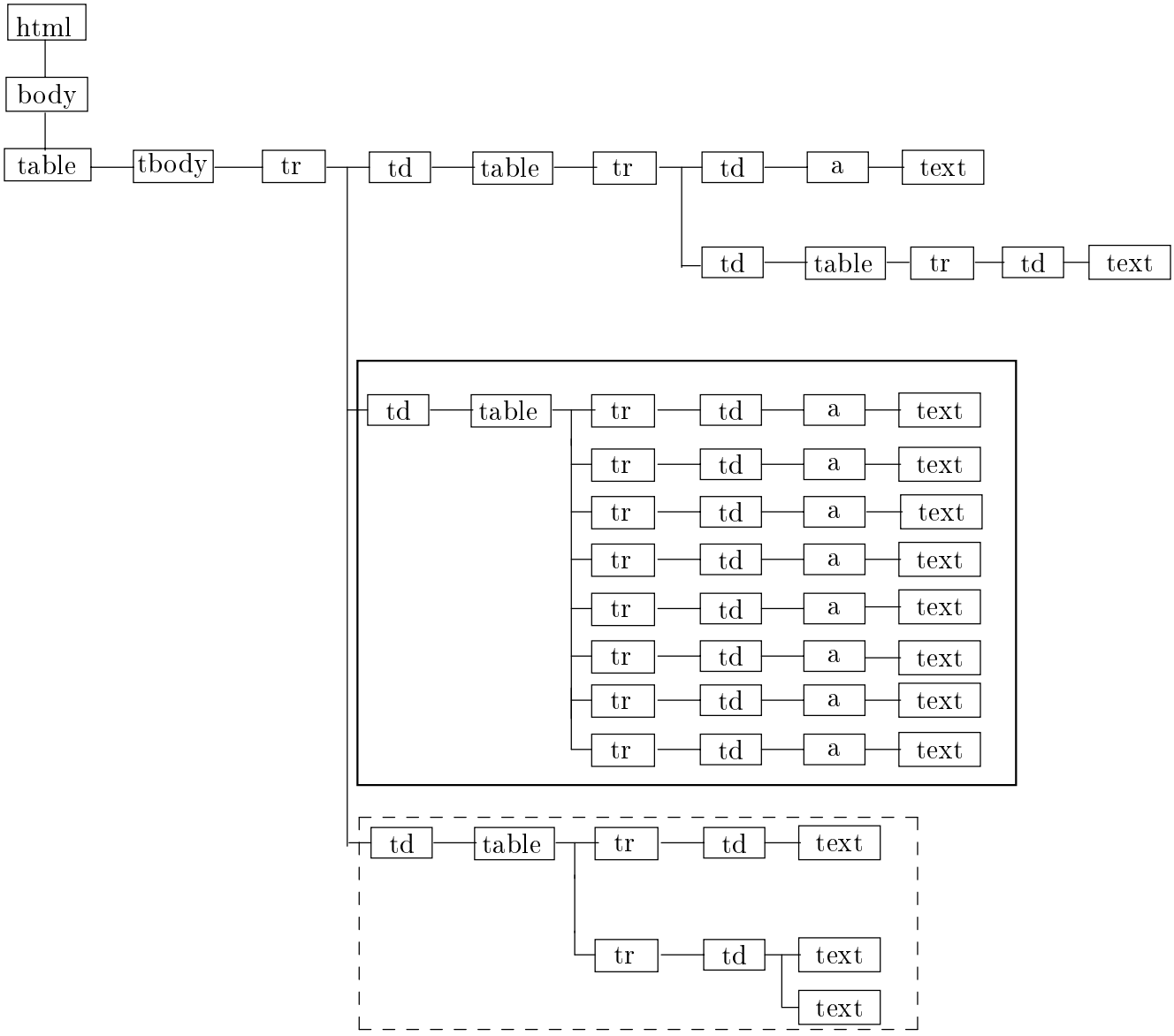

図 3.14: 図 3.13 の DOM ツリー

図 3.15: table タグで構成された例外的な領域 (テキスト平均長)

図 3.16: table タグで構成された例外的な領域 (リンクの割合)

第4章　実験

本章では、本研究で提案するウェブページにおける非コンテンツ領域の検出手法の有効性を検証するための実験について述べる。4.1 節では、提案手法の有効性を検証するために用いた実験用データについて述べる。4.2 節では、提案手法の有効性をはかる評価指標について述べる。4.3 節では、4.2 節で述べた評価指標による実験結果について述べる。4.4 節では、非コンテンツ領域を検出するために用いた素性の有効性を検証する実験について述べる。4.5 節では、非コンテンツ領域が上手く検出できていない例とその解決策について述べる。

4.1　実験データ

実験用データとして、非コンテンツ領域が付与されたウェブページの集合を用意する。しかし、全てウェブページに対して非コンテンツ領域を付与することは不可能である。そこで、ウェブディレクトリ dmoz(http://www.dmoz.jp/) に登録されているページから、ランダムサンプリングを行い実験データを取得した。

ウェブディレクトリに登録されているページの多くが TOP ページである。しかし、実験で用いるページは、TOP ページだけでなく他のページも用いたい。そこで、ウェブディレクトリ内のページをまず取得し、そのページに張られた内部リンクのウェブページを取得した。具体的には、domz に登録されているページ 46 ページをまず取得し、そのページ中の内部リンクのリンク先ページ 735 ページを取得し、合計 781 ページの実験用データを用意した。ただし、素性の選定に用いた 21 ページのウェブページ (3.3 節を参照) は実験用データに含まれていない。

筆者を含めた 4 人の大学院生がこれらのウェブページに対し非コンテンツ領域を人手で付与した。作成者には、3.1 節で述べた非コンテンツ領域の定義を説明し、ウェブページの HTML ソースに対して非コンテンツ領域を表す以下のようなコメントを追加させた。付与する際は、ウェブブラウザで表示したページと HTML ソースの両方を見ながら、非コンテンツ領域の付与を行ってもらった。

- 非コンテンツ領域の始点

  ```
  <!-- (((BEGIN NOT CONTENT -->
  ```

- 非コンテンツ領域の終点

  ```
  <!-- )))END NOT CONTENT -->
  ```

4.1.1 作成者による非コンテンツ領域の一致率

ここでは、非コンテンツ領域であると判断する場所が異なる作成者間でどの程度一致するかを調査するために、2 人 1 組で同じページに対して独立に非コンテンツ領域の付与を行った。今回は、4 人で非コンテンツ領域の付与を行ったが、2 つのページ集合を用意し、2 名の作成者が 1 つのページ集合に対して非コンテンツ領域を付与し、別の 2 名の作業がもう一方のページ集合に対して非コンテンツ領域を付与した。このときの作成者による非コンテンツ領域の一致率を求めた。一致率として『領域単位の一致率』と『テキスト単位の一致率』の 2 つの指標を用いた。

領域単位の一致率の定義を式 (4.1) に示す。

$$\text{領域単位の一致率} = \frac{2 \times A_{xy}}{A_x + A_y} \tag{4.1}$$

ここで、A_{xy} は作成者 X と Y がともに非コンテンツ領域とした領域数で、範囲とラベルが完全に一致した領域を正解とする。また、A_x は作成者 X が非コンテンツ領域とした領域数、A_y は作成者 Y が非コンテンツ領域とした領域数である

テキスト単位の一致率の定義を式 (4.2) に示す。

$$\text{テキスト単位の一致率} = \frac{2 \times L_{xy}}{L_x + L_y} \tag{4.2}$$

ここで、L_{xy} は作成者 X と Y がともに非コンテンツ領域内のテキストであるとしたラベル数である。このとき B、I の区別はせず、ともに非コンテンツ領域のテキストであるとする。

また、L_x は作成者 X が非コンテンツ領域としたラベル数、L_y は作成者 Y が非コンテンツ領域としたラベル数である。これらの一致率から、正解データ作成者の違いによる非コンテンツ領域の付与の違いがどの程度であるかを調査する。領域単位での一致率とテキスト単位での一致率の結果を表 4.1 示す。ここで、H_1, H_2, H_3, H_4 は実験データ作成者である。

表 4.1: 異なる作成者間の非コンテンツ領域の一致率

作成者	ページ数	領域単位	テキスト単位
$H_1 - H_2$	36	0.581	0.810
$H_3 - H_4$	26	0.644	0.836

- 考察
  表 4.1 から、領域単位で評価したときの一致率は作成者 $H_1 - H_2$ 間で 0.581、$H_3 - H_4$

間で 0.644 となっている。また、テキスト単位で評価したときの一致率は $H_1 - H_2$ 間で 0.810、$H_3 - H_4$ 間で 0.836 となった。これより、領域単位では約 6 割しか一致しなかったが、テキスト単位では約 8 割の一致率が得られた。この値は、作成者による非コンテンツ領域の一致率として十分に高い値とは言えない。これより、作成者によって非コンテンツ領域と認識する箇所に違いがあることがわかる。これは非コンテンツ領域の定義が曖昧で、人によって解釈が異なることが原因の一つと考えられる。したがって、非コンテンツ領域のより厳密な定義が必要である。また、後述するように、今回の実験では作成者の作業内容の理解不足が一致率を下げた要因になっていた。

- 不一致の例
  ここでは、作成者によって非コンテンツ領域の不一致が起こった例を示す。図 4.1、図 4.2、図 4.3、図 4.4 の実線で囲まれたところが、実験データを作成するときに作成者によって領域の判定に不一致が起こった所である。すなわち、2 人の内一人が非コンテンツ領域としたところを残りの一人がコンテンツ領域とした所である。

図 4.1: 不一致の例 1

図 4.1 については、この実線で囲まれた部分にはページの内容が衛星放送で放送されている、という内容が書かれている。この部分を、ページの内容に関連することが書いてあるのでコンテンツ領域と判断した人と、ページに多少関連するのだが衛星放送が直接ページとは関係ないと判断したために、判定の不一致が起こったと考えられる。

図 4.2: 不一致の例 2

図 4.3: 不一致の例 3

図 4.4: 不一致の例 4

図 4.2 については、実線で囲まれた部分は書かれていることについてさらに詳細に知りたいときにクリックするリンクである。図 4.3 の実線で囲まれた部分には、カレンダーが書かれている。図 4.4 の実線で囲まれた部分は、ページ作成者が記載した店の住所や電話番号などが書かれている。

図 4.2、図 4.3 の例は、情報検索の立場から考えて非コンテンツ領域と判定するのが一般的であると思われる。また、図 4.4 の例は、ウェブページの情報として重要なものであるのでコンテンツ領域と判定するのが自然である。これらの不一致は作成者の作業内容の理解不足が原因である。したがって、作業前に非コンテンツ領域についての説明を徹底すれば一致率はもう少し高くなったと思われる。

4.2 評価指標

提案システムの評価指標を以下に述べる。

- ラベルの正解率

$$L_{nc} = \frac{\text{正解データとシステム出力のラベルの一致数}}{\text{正解データのラベル数}} \tag{4.3}$$

L_{nc} は、システムが出力したラベル (B,I,O) がどれぐらい正解しているかを表したものである。

- 非コンテンツ領域検出の領域単位での再現率と精度

$$R_r = \frac{\text{正解データとシステムが出力した非コンテンツ領域の一致数}}{\text{正解データの非コンテンツ領域数}} \tag{4.4}$$

$$R_p = \frac{\text{正解データとシステムが出力した非コンテンツ領域の一致数}}{\text{システムが出力した非コンテンツ領域数}} \tag{4.5}$$

ここでは、非コンテンツ領域の領域範囲とラベル (B,I) が完全に一致しているかの評価を行っている。R_r は、非コンテンツ領域の検出を領域単位でみた場合の再現率を表し、R_p は、非コンテンツ領域の検出を領域単位でみた場合の精度を表している。

- 非コンテンツ領域検出の領域単位での F 値 (F-measure)

$$R_F = \frac{2R_rR_p}{R_r + R_p} \tag{4.6}$$

これは、非コンテンツ領域が領域単位でどれくらい上手く検出できているかを示す評価指標である。

- 非コンテンツ領域検出のテキスト単位での再現率と精度

$$B_r = \frac{\text{正解データとシステムの出力の両方で非コンテンツ領域にあるテキスト数}}{\text{正解データにおける非コンテンツ領域内のテキスト数}} \tag{4.7}$$

$$B_p = \frac{\text{正解データとシステムの出力の両方で非コンテンツ領域にあるテキスト数}}{\text{システムが非コンテンツ領域と出力したテキスト数}} \tag{4.8}$$

ここでは、非コンテンツ領域内のテキストのラベルに対して、正解ラベルとシステムが出力したラベルがどれぐらい一致しているかを評価している。B_r は、非コンテンツ領域の検出をテキスト単位で評価した場合の再現率である。そして、B_p は、非コンテンツ領域の検出をテキスト単位で評価した場合の精度を表す。ここで、ラベルが B か I かの区別は行っていない。

- 非コンテンツ領域検出のテキスト単位での F 値

$$B_F = \frac{2B_rB_p}{B_r + B_p} \tag{4.9}$$

これは、非コンテンツ領域がテキスト単位でどれくらい上手く検出できているかを示す評価指標である。

- コンテンツ領域の誤判定率

$$FP_c = \frac{\text{誤って非コンテンツ領域と判定されたテキスト数}}{\text{正解データにおけるコンテンツ領域のテキスト数}} \tag{4.10}$$

FP_c は、コンテンツ領域と判定されるべきテキスト (正解データでラベルが O となっているテキスト) が、システムによって非コンテンツ領域であると判定されたテキスト (B または I と判定されたテキスト) の割合である。

4.3 実験結果

3.3 節で述べた素性を用い、学習データからチャンキングのモデルを学習し、ウェブページにおける非コンテンツ領域の検出を行った。実験方法としては、実験用データを 5 分割して、1 つをテストデータ、残りを学習データとして実験する。それを 5 回繰り返す 5 分割交差検定 (5-fold cross validation) を行った。提案手法との比較のために、ベースラインとして常に O ラベル (コンテンツ領域) を出力するシステムのラベルの正解率 L_{bl} を求めた。その結果を表 4.2 に示す。

表 4.2: 提案手法の実験結果

L_{bl}	L_{nc}	R_r	R_p	R_F	B_r	B_p	B_F	FP_c
0.698	0.769	0.135	0.296	0.185	0.431	0.694	0.532	0.0694

表 4.2 からシステムがテキストについて出力したラベル (B,I,O) の正解率 L_{nc} は 0.769 で、ベースライン L_{bl} の 0.698 を大きく上回ることがわかる。非コンテンツ領域検出における領域単位での再現率 R_r は 0.135 で、精度 R_p は 0.296 となっている。また、非コンテンツ領域検出におけるテキスト単位での再現率 B_r は 0.431 で、精度 B_p は 0.694 となっている。非コンテンツ領域の領域単位での精度 R_p は約 3 割と低い。一方、非コンテンツ領域内のテキスト単位での精度 B_p は約 7 割であった。この結果は、非コンテンツ領域の範囲とラベルを完全に検出することは難しいが、非コンテンツ領域を部分的に検出することがある程度できていることを示している。しかし、まだ十分なものとは言えないのでシステムの改良が必要である。また、コンテンツ領域におけるテキストのラベルを誤って B もしくは I と判定してしまった割合 FP_c は 0.0694 となっている。約 7%程度しか情報として有用なコンテンツ領域が失われていないことが分かる。

4.4 素性の評価

チャンキングに用いた素性の有効性を検証するために以下の実験を行った。まず、表 4.3 に挙げる素性について、それを素性として用いないでチャンキングのモデルを学習し、

非コンテンツ領域の検出を行った。実験方法は 4.3 節と同じように 5 分割交差検定を行った。その結果を 4.3 節の評価指標で評価した。結果を表 4.4 に示す。

表 4.3: 有効性を検証する素性

	有効性を検証する素性
case1	非コンテンツ領域によく現れるキーワード
case2	動詞、形容詞が含まれているかどうか
case3	テキスト長
case4	<table>タグ内の平均テキスト長、リンクの割合
case5	DOM ツリーの末端が<a>タグか？
case6	DOM ツリーにおけるテキストの直前の 3 つのタグ
case7	直前のテキストと比較した DOM パスの深さの変化

表 4.4: 素性の有効性検証

	L_{nc}	R_r	R_p	R_F	B_r	B_p	B_F	FP_c
case1	0.725	0.0607	0.151	**0.0866**	0.284	0.593	**0.384**	0.0716
case2	0.730	0.129	0.220	0.163	0.421	0.580	0.488	0.120
case3	**0.700**	0.149	0.119	0.132	0.493	0.517	0.505	**0.188**
case4	0.730	0.118	0.135	0.126	0.324	0.598	0.420	0.0815
case5	0.767	0.145	0.292	0.194	0.408	0.705	0.517	0.0609
case6	0.733	0.121	0.344	0.179	0.340	0.597	0.433	0.0867
case7	*0.771*	0.156	0.274	*0.199*	0.514	0.675	*0.584*	0.0943

ここで R_F と B_F に着目すると、これらの値が最も低かったのが case1 のときであった。これは、非コンテンツ領域の検出の正確さを重視したとき、最も有効な素性が非コンテンツ領域を示唆するキーワードであることがわかる。一方、L_{nc} と FP_c に着目すると、これらの値が最も低かったのが case3 のときであった。これは、コンテンツ領域の誤検出の少なさと非コンテンツ領域の検出の正確さの両方を重視したとき、最も有用な素性がテキスト長であるということがわかる。一方、case7 の L_{nc} と R_F と B_F を見ると、提案システムの結果よりも良い値になっている。これは、case7 の素性が原因で、非コンテンツ領域の検出の正解率が下がった可能性があることを表している。したかって、『直前のテキストと比較した DOM パスの深さの変化』という素性はそれほど有効ではなかったことがわかる。

次に、実験で用いた素性『非コンテンツ領域によく現れるキーワード』の自動選別がどれぐらい上手くできているかの評価を行う。比較実験として、素性として用いるキーワードを人手で決定し、素性とする。具体的には表 4.5 に示す 6 つのキーワードの有無を素性

として用いた。これらのキーワードは、3.3 節で述べた 21 個のウェブページを用いた有効な素性の調査の過程で発見したものである。実験を行う条件としては、キーワードの選別以外は 4.3 節で行った実験と同じ条件で行った。実験結果を表 4.6 に示す。

表 4.5: 人手で設定した非コンテンツ領域を示唆するキーワード
目次、検索、Link、広告、Copyright、著作

表 4.6: 固定のキーワードを用いた場合の実験結果

L_{nc}	R_r	R_p	R_F	B_r	B_p	B_F	FP_c
0.732	0.0989	0.212	0.135	0.354	0.604	0.412	0.0878

人手で選択したキーワードを用いた場合の実験結果の表 4.6 とキーワードの自動選別を行った提案手法の結果の表 4.2 を比較すると、提案手法の方が良い値を得た。このことから、キーワードの選別がうまく行われていることがわかった。

自動選別されたキーワード表 4.7 に示す。これらのキーワードは実験に用いたデータ全体から選別したものである。実験を行うときには、学習データからキーワードを自動選別しているので、表 4.7 のキーワード全てを素性として使用しているわけではない。表 4.7

表 4.7: 自動選別されたキーワード

C、Co、ホーム、c、情、Copyright、TOP、All、トップ、マップ、Reserved、HOME、Rights、基、プライバシー、ニュース、ポリシー、reserved、Adobe、Reader、宮、rights、住、ＴＯＰ、京、アンケート、史、Ltd、Inc、規約、買い物、先頭、取扱 、食材、ENGLISH、フォト、見積もり、カテゴリー、ペット、キッズ、表記、ギャラリー、Player、初、INDEX、more、Simulingual

をみると、自動選別したキーワードのすべてが非コンテンツ領域を一般的に示唆するようなものではないが、『ホーム』や『Copyright』、『TOP』など非コンテンツ領域によく現れるキーワードと思われるものが取り出せていることがわかる。

4.5　エラー分析

今回の実験結果から、非コンテンツ領域の検出が上手く行えていない実例をいくつか下に示し、その解決策について示す。

検出エラー 1

図 4.5 の実線で囲まれた部分は正解データでは非コンテンツ領域であるとなっているのだが『｜ ＩＮＤＥＸ｜車輌 ｜フィルタ｜インテリア｜工業用途｜生活資材｜リサイクル・環境｜ ＡＮＥＸ２００６｜』の中の『INDEX』だけがシステムによってコンテンツ領域のテキストであると判断されてしまっている。この原因として考えられるのは、<table>タグ内にあるテキスト『INDEX』だけがアンカー文字列でないことが誤りの原因であると思われる。同じ<table>タグ内にある要素である『車輌』や『フィルタ』などが非コンテンツであると判定されているので、『INDEX』が非コンテンツと判定できそうである。そのために、<table>内のテキストのリンクの割合を素性として用いている。しかし、今回の実験ではページの先頭からテキストのラベルを順番に判定しているため、誤ってコンテンツ領域と判定してたと思われる。

そこで、チャンキングに用いる素性として判定するテキストの後ろにある要素の情報を考慮することによって、今回の誤った判定を正すことができるのではないかと考える。

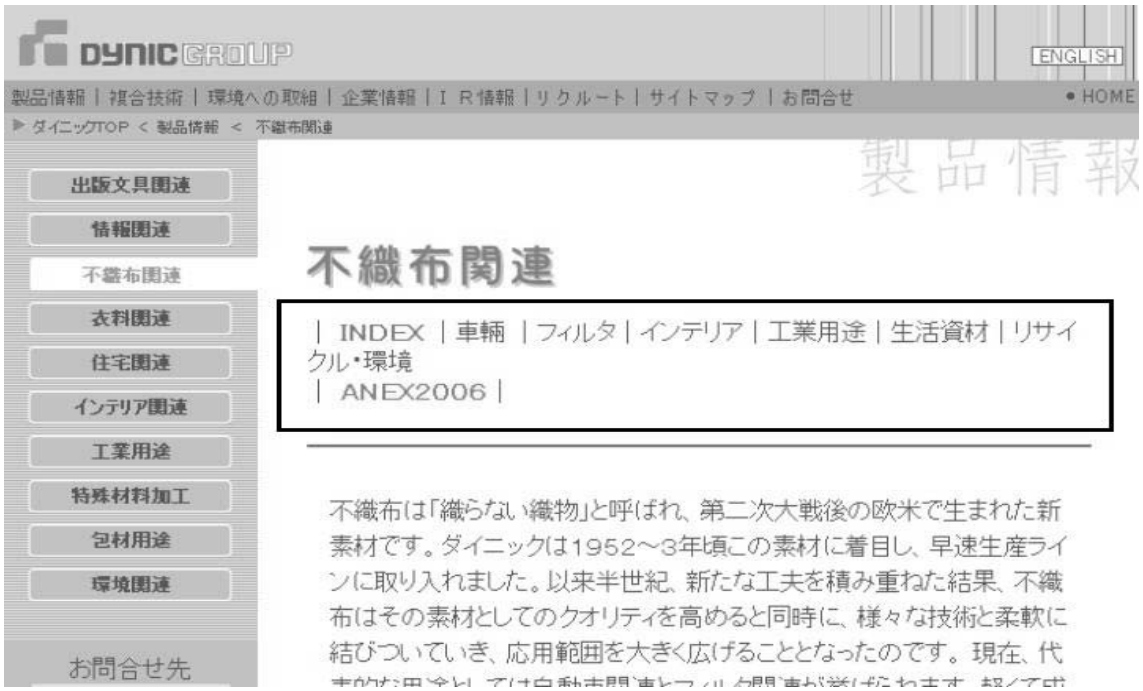

図 4.5: エラー解析 1

検出エラー 2

図 4.6 のページにおいて、実線で囲まれた領域　はページのナビゲーション目的のリンクグループであるので非コンテンツ領域と、領域　はページの本文になっているのでコンテンツ領域と判定されてほしい。しかし、実験の結果では領域　、領域　の両方とも非コンテンツ領域と判定されてしまっている。この原因として、領域　と領域　の要素の素性がほとんど同じものになってしまっていることが挙げられる。両者に対して同じ素性になっているものは、テキストの上位 3 つのタグ、直前のテキストと比較した DOM ツリーの深さの変化である。図 4.6 を DOM ツリーで表した図 4.7 を見るとわかるように、テキスト『INDEX』や『関連サイト』についての上位 3 つのタグが table-tr-td となっているのと同様に『ファッションの・・・』についての上位 3 つのタグも table-tr-td となっている。また、DOM ツリー (図 4.7) におけるテキスト『ファッションの・・・』が位置する深さは直前のテキスト『 | 』と比較しても変化がない。このために、領域　と領域　は同じ<table>内に存在すると判断され、2 つの table が違うものであることが認識できていない。すなわち、図 4.7 の DOM ツリーと 4.8 の DOM ツリーからは、各テキストに対して、DOM ツリー上の HTML や DOM ツリー上の深さの変化について全く同じ素性が得られる。このことから、領域　は領域　と同じ非コンテンツであるとシステムは判定している。このように、HTML の構造が変化が認識されず同じ構造内にあるものとされてしまっているものが存在した。これでは、ウェブページ構造の情報として有用な部分が失われてしまうことになる。

このような問題を解決するために、直前のテキストが DOM ツリーにおいて位置する深さを素性にするのではなく、直前のテキストとの同一ノードまでの距離を素性とすれば、構造の変化の情報を素性として用いることができるのではないかと思われる。図 4.7 を見ると、テキスト『 | 』の直前のテキスト『関連サイト』の同一ノードとなる<td> までの距離は 1 となる。一方、テキスト『ファッションの・・・』の直前のテキスト『 | 』の同一ノードは最左にある<td>となり、このとき距離は 4 となる。このように、ルートからテキストまでの深さの変化を素性とするのではなく、直前のテキストとの共通上位ノードまでの距離を素性とすることで HTML の構造変化の情報をモデルに反映させることができると考えられる。

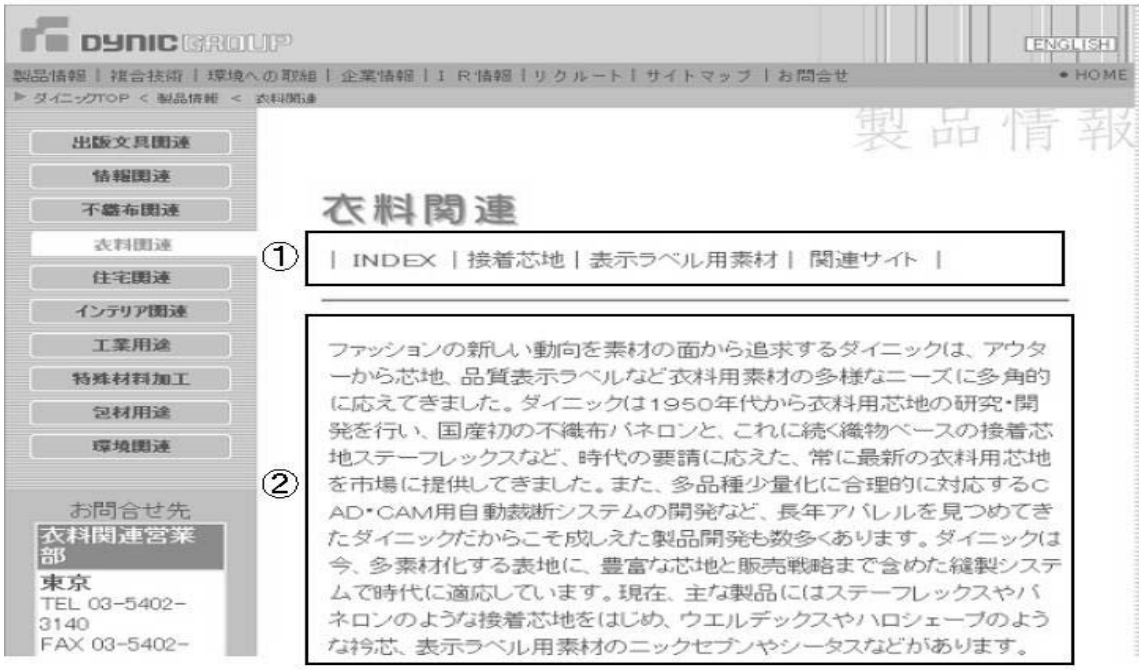

図 4.6: エラー解析 2

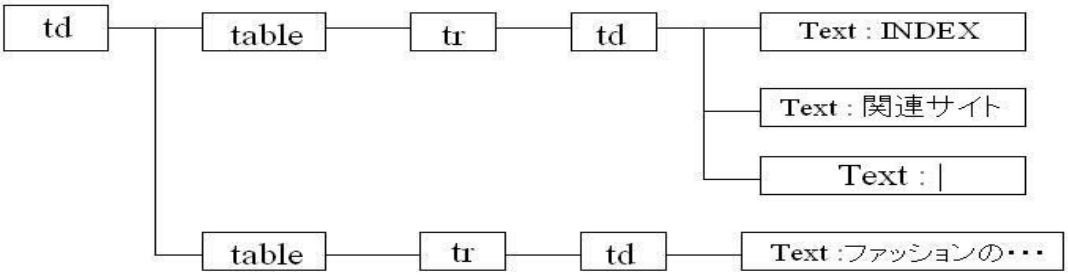

図 4.7: 図 4.6 の DOM ツリー

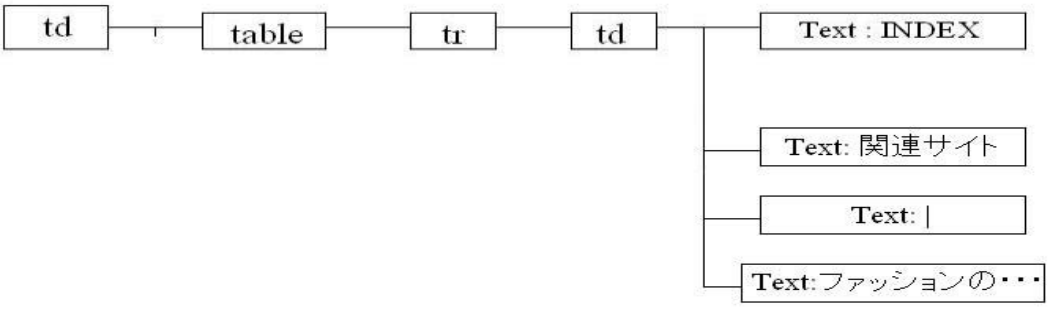

図 4.8: 誤認識した DOM ツリー

第5章　おわりに

5.1　まとめ

本研究では、ウェブページにおいて情報として有用なコンテンツ領域と情報として有用でない非コンテンツ領域が存在し、そのページの非コンテンツ領域が情報検索やウェブマイニングなどのアプリケーションに対して悪影響を及ぼすことを問題としている。その問題を解決するため、チャンキングによってウェブページにおける非コンテンツ領域を検出する手法を実装した。チャンカーとしてYamChaを用いてウェブページのHTMLタグによって分割されたテキストに対し、コンテンツ領域か非コンテンツ領域かのラベル付けを行うことで領域の検出を行う。

ウェブページにおける非コンテンツ領域の検出手法の有効性を確認するために、実験に用いるデータを次のようにして取得した。まず、ウェブディレクトリ内のウェブページを46ページ取得し、この46ページのページ内に張られた内部リンクの735ページを取得し、合計781ページを実験用データとした。このデータを5分割交差検定によって学習とテストを繰り返し、提案手法の有効性を確認した。

提案システムのラベルの正解率の評価指標としてL_{nc}、また、非コンテンツ領域検出における領域単位の評価指標としてR_r, R_p、テキスト単位の評価指標としてB_r, B_pを用いた。ラベルの正解率L_{nc}は0.769となり、システムが出力したテキストに対するラベルが全てOであるとした場合のベースラインL_{bl}は0.698であるので、この値を大きく上回っていることがわかった。そして非コンテンツ領域検出における領域単位の精度R_pは約3割、非コンテンツ領域検出におけるテキスト単位の精度B_pは約7割であることがわかった。このことから、提案システムが、非コンテンツ領域の範囲とラベルを完全に検出することは難しいが領域を部分的に検出することができているのが分かる。しかし、R_pもB_pも十分高いと言えないので、更なる手法の改良を行う必要がある。また、情報として有用なコンテンツ領域が誤って非コンテンツ領域と判定されていないかの評価指標FP_cは0.0694となって、ウェブページにおける有用な情報を大きく失っていないことがわかった。

次に、チャンキングに用いた素性の有効性を検証するために、検証したい素性を用いないで学習したモデルとの比較実験を行った。非コンテンツ領域の検出のみを重視した場合に、最も有効な素性であると確認できたのは『非コンテンツ領域によく現れるキーワード』であった。また、コンテンツ領域の誤検出の少なさと非コンテンツ領域の検出の両方を重視した場合に、最も有効な素性であると確認できたのは『テキスト長』であった。一方、素性『直前のテキストと比較したDOMパスの深さの変化』を用いないで実験を行っ

た場合は、提案手法よりうまくラベルの判定が行えている場合もあった。これから、この素性があまり有効でないことがわかる。

また、チャンキングに用いる素性のキーワードの自動選別がどれぐらいうまくいっているかの実験を行った。実験に用いるキーワードは、事前に採集したページを調査し選定したキーワードを用いた。この実験結果とすべての評価指標において自動選別されたキーワードを用いた実験結果を比較すると、自動選別されたキーワードを用いたほうが良い結果が得られた。この結果から、キーワードの自動選別が上手く行われていることがわかった。

5.2 今後の課題

本研究で作成したウェブページにおける非コンテンツ領域を検出するシステムは、実験結果からもわかるように十分高い精度で非コンテンツ領域を検出できていない。そこで、提案システムの改良するべきところをいくつか挙げる。今回非コンテンツ領域の検出に用いた素性は、事前に採集した 21 ページのウェブページを調査し選定した。しかし、これらの素性がチャンキングの素性として十分であるとは限らない。このことから、様々なウェブページを調査し、非コンテンツ領域を検出する素性として有用なものを見つけることで非コンテンツ領域の検出精度が改善するのではないかと考えられる。例えば、本論文では用いていないが、テキストに『、』や『。』があるかどうかも素性として有用であると考えることができる。テキストに『、』や『。』が含まれていれば、そのテキストが文を構成している可能性が高いので、コンテンツ領域と非コンテンツ領域を判別する素性として有用であるかもしれない。

また今回、<table>タグ内のテキストの平均長やアンカーテキストが存在する割合などがナビゲーション目的の目次などの判定に有効であると考え、チャンキングの素性として実験に用いている。このことから、ナビーゲーション目的の目次やリンクグループなどを構成することが多いリストタグ<ol>,<ul>,<dl>においても<table>タグと同様にリストタグ内のテキストの平均長、アンカーテキストの割合を素性として加えることでよりよい結果を得ることができるのではないかと思われる。

また、今回チャンキングに用いた素性の与え方を改良することで実験結果を向上させることができると思われる。まずテキスト長について、本手法でこの素性値は表 3.1 のように与えている。しかし、この素性値の与え方が必ずしも最も良い与え方であるとは限らない。同様に、表 3.4 のように与えられた<table>タグ内のテキストの平均長と表 3.5 のように与えられた<table>タグ内のリンクの割合も素性値の定義を見直す必要がある。これらの素性値の与え方として、最適な値のセットを求めることでシステムの改良をはかることができると思われる。

本論文では、アプリケーションとして情報検索を想定し、ウェブページから非コンテンツ領域を検出し、その領域中の単語を索引語としないことで情報検索の向上を見込んでいる。しかし、今回の実験では提案システムがどれぐらい情報検索の向上に貢献しているか

の評価は行っていない。よって、提案システムによる非コンテンツ領域の検出の精度を上げ、次の段階として非コンテンツ領域を検出し、その領域内の単語を無視することで情報検索の精度をどれぐらい向上できるかを評価する必要があると思われる。

謝辞

北陸先端科学技術大学院大学・自然言語処理学講座の島津明先生、白井清昭先生には様々な御指導していただき、心から感謝しております。また、中村誠助手、同期の皆様にも助けていただき本当にお世話になりました。特に、実験に用いるデータの作成をしていただいた竹島正泰氏、九岡佑介氏、長内亘氏には深く感謝しています。

お世話になった皆様に心から御礼を申し上げます。

参考文献

[1] Shian-Hua Lin,Jan-Ming Ho,Discovering Informative Content Blocks from Web, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,pp.588-593,2002

[2] Yu, Shipeng and Cai, Deng and Wen, Ji-Rong and Ma, Wei-Ying, Improving Pseudo-Relevance Feedback In Web Information Retrival Using Web Page Segmentation, In Proceedings International WWW Conference, Budapest, Hungary, 2003

[3] 浅原正幸, 松本裕治, 形態素解析とチャンキングの組み合わせによる日本語テキスト中の未知語出現箇所同定, 『情報処理学会研究報告』2003-NL-154, pp47-54, 2003

[4] Sandip Debnath, Prasenjit Mitra, C. Lee Giles, Identifying Content Blocks from Web Documents, ISMIS, 2005

[5] Yudong Yang, Hongjiang Zhang, HTML Page Analysis Based on Visual Cues, ICDAR, 2001

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

[7] Taku Kudo, Yuji Matsumoto, Chunkin with Support Vector Machines, NAACL, 2001

[8] 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.

[9] 南野朋之, 斉藤豪, 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.