Web ページにおけるナビゲーション領域検出を利用した非主要部分特定手法

宇田 賢広† 松本 章代†† 小西 達裕† 高木 朗††† 小山 照夫†††† 三宅 芳雄††††† 伊東 幸宏†

†静岡大学 〒432-8011 静岡県浜松市城北 3-5-1
††東北学院大学 〒981-3193 宮城県仙台市泉区天神沢 2-1-1
†††言語情報処理研究所 〒192-0919 東京都八王子市七国 3-1-23
††††国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
†††††放送大学 〒261-8586 千葉県千葉市美浜区若葉 2-11
E-mail: riir@inf.shizuoka.ac.jp

あらまし ニュースサイトや EC サイト，ブログページなど Web ページには様々な情報が含まれているが，多くの Web ページにはそのページの主題とは直接関係ない部分（非主要部分）が含まれている．先行研究では，Web ページの構造的な切れ目を用いて Web ページをブロック化し，教師データを用いて非主要部分と主要部分を決定木学習させることで Web ページの非主要部分を特定する手法を提案した．本稿では，Web ページ中のナビゲーション領域を検出する手法を提案し，これを用いて先行研究の Web ページのブロック化を見直し，さらに特定のナビゲーションと主要部分との位置関係の典型的パターンを手掛かりに Web ページの非主要分部分の特定精度を向上させる．

キーワード Web ページ構造解析，ナビゲーション領域，非主要部分特定

1. はじめに

近年，Web 上の情報量は爆発的な増加傾向にあり，それに伴い Web ページ中の情報を有効利用するための情報検索や Web マイニング技術は不可欠となっている．しかし，多くの Web ページには，サイト制作者が閲覧者に伝えたい情報とは直接関係のない情報が含まれている領域が存在する．たとえば，Web ページ中に存在するナビゲーションメニュー，広告，アクセスランキング，関連記事などといったような領域のことである．こういったサイト制作者が閲覧者に伝えたい情報とは直接関係のない情報が含む領域（以下，非主要部分）を排除する，もしくはサイト制作者が閲覧者に伝えたい情報を含む主要部分のみを抽出できれば，WWW 検索エンジン，携帯端末向けの Web ページ変換，コンテンツフィルタリングなどのシステムへの応用が期待できる．

先行研究[1]では，服部ら[2]が提案した HTML のタグの深さに基づくコンテンツ間距離を利用した Web ページのブロック化を行い，各ブロックの広告やナビゲーションメニューといった非主要部分とそれ以外の主要部分の教師データを用いて，各ブロックのテキストのバイト数，言語的性質（文を構成する形態素の数など）およびリンクの有無や自サイトへのリンクであるかという特徴情報をもとに決定木学習させることで Web ページの非主要部分を特定する手法を提案した．しかし，先行研究の手法では，非主要部分の特定精度が，Web ページのブロック化の精度に依存するため，正しくブロック化されずに主要部分と非主要部分が混在する場合に対しては正しく非主要部分を特定できない．また，非主要部分をナビゲーションメニューや広告といった情報ごとに区分しているため，出現頻度が低い非主要部分については，決定木学習のための教師データが少ないことから特定精度が悪いという問題があるが，教師データを増やすにはコストがかかる．

そこで本研究では，Web ページ中のナビゲーション領域に着目する．ナビゲーション領域は，ページ中のヘッダ領域やフッタ領域，サイド領域といった主要部分の前後左右の領域に出現しやすいため，主要部分のブロックとの区切りにすることでより正しいブロック化を行うことができるはずである．また，特定のナビゲーションは，主要部分との位置関係が固定される傾向にある，または主要部分自身へのナビゲーションであるため，これらを典型的なパターンとして非主要部分または主要部分を同定する手掛かりに利用することで特定精度を向上させる．

2. 関連研究

Web ページの主要部分または非主要部分の抽出に関する研究は多く存在する．ここでは，Web ページ群を

対象とした手法と単一の Web ページを対象とした手法について述べる．

Web ページ群を対象とした手法としては，ある Web ページの主要部分は，他の Web ページに出現しない傾向にあることを利用し，Web ページを大量に収集し，Web サイトにおいて複数回出現する DOM 部分木などのタグパターンや文字列部分を非主要部分と見なし，削除することで主要部分を抽出する手法が多く提案されている．Lin ら[3]は，同じサイト内の Web ページを収集し，Web ページを Table タグに基づいたブロックに分割し，各ブロックに対してキーワード抽出を行い，サイト内でのエントロピーを計算する．計算されたエントロピーの値が閾値以下であれば，非主要領域としている．吉田ら[4]は，W3C（World Wide Web Consortium）で定義されたブロックレベル要素に支配されたテキストや画像を最小のブロックとして分割し，ブロックが持つタグ数やテキスト数を特徴情報としてブロック間のコサイン類似度を計算する．類似度が閾値よりも高いブロックを非主要部分として特定する手法を提案している．これらは Web ページをブロックに分割し，コンテンツを特定するという点で本研究と関連が深い．

単一の Web ページを対象とした手法としては，人手によって重要度や主要，非主要などの正解が付与された教師データを用いて，機械学習を行い，学習結果に基づき非主要部分，または主要部分を特定する手法が提案されている．鶴田ら[5]は，主要な DOM ノード情報が付与された教師データを用いて，一般的な Web ページにおいて，ブラウザ上のどの位置に主要 DOM ノードが出現するかを学習することによって，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．ただし，この手法は平均的な 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持つ．そこで鶴田らは，上記の手法に加え，レイアウトが類似する 2 つの Web ページにおいて，主要部分は，それらの間で類似する領域に存在するという仮定に基づき，教師データ中の DOM ノードが持つレイアウト情報を用いて，対象となる Web ページとの類似度を計算することによって，主要な DOM ノードを抽出する手法を提案し，既存手法との組み合わせによる評価を行っている[6].

ここで上述のようにナビゲーション領域はページの構造か主要・非主要部分を特定するのに有用な情報を持っていると考えられる．しかし，これらの研究ではナビゲーションに着目した解析は行っていない．

また，非主要部分の特定技術は，様々な Web アプリケーションの前処理に利用できると考えられる．特に本研究では，非主要部分をナビゲーション領域とそれ以外という使い方が可能であるため，非主要部分の扱いをアプリケーションやユーザの求めによって，柔軟に対応できる．たとえば，Web 検索エンジンにとっては，広告やナビゲーション領域などのページの主要部分以外はすべて非主要部分であると考えられる．一方，Web ページを携帯端末などの小さい画面に表示するためにページの主題だけを抽出することが目的の場合においては，すべてのナビゲーション領域を排除すればいいわけではなく，ユーザが必要に応じて，ナビゲーション領域を画面に表示したい場合が存在する．たとえば EC サイトにおける“ログイン”や“買い物かごに入れる”といったナビゲーションは常時表示することが望まれることが考えられる．

3. ナビゲーション領域について

本稿では，Web ページに出現するナビゲーション領域にはどういったものがあるのかを検討し，検出する必要のあるナビゲーションについて検討するとともに，各ナビゲーションの検出処理について述べる．

3.1 典型的なナビゲーション領域

我々の研究では文献[7]を参考に Web ページにおける典型的なナビゲーション領域を以下のように分類した．

- カテゴリナビゲーション：

サイトの主要な第 1 階層カテゴリへの移動，または同一階層や下層のカテゴリへの移動を可能にするナビゲーション．

- パンくずリストナビゲーション：

主に現ページの階層位置を表現したもので，現ページからトップページまでの各階層コンテンツへの移動を可能にするナビゲーション．

- ページングナビゲーション：

ページ移動を誘導するためのナビゲーション．文脈に沿った誘導，動的に生成されたページへの誘導，時系列に沿った誘導などがある．

- サイト情報ナビゲーション：

サイト自体の情報へアクセスするためのナビゲーション．サイトマップ，プライバシーポリシー，個人情報保護方針，運営者情報などの典型的な例である．

- 関連情報ナビゲーション：

サイト制作者が一つの観点からまとめたコンテンツへのナビゲーション．新着，認識，おすすめ記事などが典型的な例である．

- ユーティリティナビゲーション：

買い物かご，ログイン，My ページ，サイト内検索などのサイトが持つ機能やツールを利用する際に使用するためのナビゲーション．

- ページ内ナビゲーション：

ページ内の上部やコンテンツなどへアクセスするた

めに配置されるナビゲーション.

3.2 検出処理を行うナビゲーション領域

3.1 節で定義した典型的なナビゲーション領域のうち，カテゴリナビゲーションについては先行研究で区分されたナビゲーションメニューのことであり，検出についても先行研究での決定木による特定を信頼することにする．よって，3.2 章および 3.3 章ではカテゴリナビゲーション以外の典型的なナビゲーション領域について述べる．

また本研究では，後述するクローズドテスト用の Web ページデータ 50 ページを対象とする事前調査を行い，典型的なナビゲーションが出現する位置傾向，タグパターン，リンク文字列，タグ属性値，リンク先アドレスといった特徴情報からの検出処理について検討した．

その結果，関連情報ナビゲーションについては，ページレイアウトによって，出現位置がかなり左右されてしまうことが分かった．またサイトのトップページなどは内部階層にあるコンテンツへのナビゲーション自体がそのページにおける主要部分となる傾向にあり，特に関連情報ナビゲーションは主要部分に見えるケースが多く存在した．そのため，関連情報ナビゲーションについては，検出しても主要部分及び非主要部分の同定手法に応用することは難しいため，検出しない．

以上のことから，3.3 節で述べる検出処理を行うナビゲーション領域は,パンくずリストナビゲーション，ページングナビゲーション，サイト情報ナビゲーション，ユーティリティナビゲーション，ページ内ナビゲーションとする．

3.3 典型的なナビゲーション領域の検出処理

3.2 節で前述した通り，クローズドテスト用の Web ページデータ 50 ページを対象とする事前調査を行い，典型的なナビゲーションの出現事例からタグパターン，リンク文字列，タグ属性値，リンク先アドレスといった形態上と HTML タグ上の特徴から各ナビゲーションの種類と特徴情報を(A)と(B)にまとめた．また，検出処理については,クローズドデータにおける(B)の特徴情報の有無をもとに一番よりよく検出できるアルゴリズムを検討し，(C)にまとめた．

① パンくずリストナビゲーション

(A) 検出ナビゲーション種類

A-1．リンク，及び区切り記号“＞”で構成

A-2．リンク，区切り画像で構成

(B) 特徴情報

B-1．id または class 属性にパンくずリストを示す文字列を含む（breadcrum, topicpath, dirnavi, pannavi）

B-2．パンくずリストナビゲーション中の先頭に特徴的な非リンク文字列を含む（現在位置:, ThisPage:）

B-3．リンク文字列，区切り記号“＞”が 1 個以上で構成されたタグパターンを含む

B-4．リンク文字列，区切り記号画像（alt 属性値が“の中の”）が 1 個以上で構成されたタグパターンを含む

B-5．箇条書きタグ（ol,ul）中の各項目のすべてがリンク文字列または最後の項目のみ非リンク文字列のタグパターンを含む

B-6．各項目リンク先アドレスの階層が深くなっているタグパターンを含む

(C) 検出アルゴリズム

Step-1．B-1 が True なら Step6 へ

Step-2．B-2 が True なら Step6 へ

Step-3．B-3 かつ B-6 が True なら Step6 へ

Step-4．B-4 が True なら Step6 へ

Step-5．B-5 かつ B-6 が True なら Step6 へ

Step-6．DOM ツリーからそのコンテンツ部分を支配するタグを特定し，パンくずリストナビ領域を検出する．

② ページングナビゲーション

(A) 検出ナビゲーション種類

A-1．数字リンク，ページ遷移を示すリンク文字列（下記の B-2）または記号の組み合わせ

A-2．ブログ型（<<記事タイトル　記事タイトル>>）

(B) 特徴情報

B-1．id または class 属性にパンくずリストを示す文字列を含む（pager，pagenavi，paging，pagenum）

B-2．ページ遷移を示すリンク文字列または画像タグの alt 属性値を含む（数字リンク，次，前，NEXT, PREV 等々）

B-3．先頭にページ遷移を示す記号を含むリンク文字列または末尾にページ遷移を示す記号（“<<”や“>>”）を利用したリンク文字列を含む（ただし，上記のページ遷移を示す記号が先頭と末尾に存在する場合は，リンク見出しを強調するための記号とみなす）

B-4．昇順に並べられた数字リンクを含む（ただし現ページの数字は非リンクであるため，非数字リンクが 1 つ存在することを許容）

B-5．現ページ URL と同階層の別ページへのリンク先アドレスを持つリンクで構成されたタグパターンを含む

(C) 検出アルゴリズム

Step-1．B-1 が True なら Step5 へ

Step-2．B-2 が True なら Step5 へ

Step-3．B-3 かつ B-5 が True なら Step5 へ

Step-4．B-4 かつ B-5 が True なら Step5 へ

Step-5．DOM ツリーからそのコンテンツ部分を支配するタグを特定し，ページングナビ領域を検出する．Step-4 から検出した領域は A-2 のページングナビであり，それ以外は A-1 のページングナビとする．

③　サイト情報ナビゲーション

サイト情報ナビゲーションは，サイトに関する情報にアクセスするためのナビゲーションであり，コンテンツに対するナビゲーションではないため，どの Web サイトにおいても共通に出現しうるナビゲーションである．しかし，ナビゲーションの見た目や構造はサイト制作者ごとに違ってものとなるため，パンくずリストナビゲーションやページングナビゲーションの様な特有の形態上，および HTML タグ上の特徴は持たない．

そこで，5.1 節で後述する収集した 1180 ページの Web ページに出現したリンク文字列群の中から頻出する名詞キーワードからサイトに関する情報にアクセスできると推定できるキーワードを目視で確認した．この Web ページは，2 つのキーワードの組み合わせによる検索結果の上位 100 ページであるため，あるキーワードの組み合わせで必ずでてくるリンク文字列が頻出してしまうケースが想定できることから，最低 200 ページ以上に出現するリンク文字列を抽出し，典型的なサイト情報ナビゲーションを示すリンク文字列と同じとして B-1 の 33 個のキーワードを得た．そして，クローズドテスト用の Web ページ 50 ページ中の主要部分に上記のリンク文字列が出現しないことを確認した．

(A) 検出ナビゲーション種類

A-1．リンク文字列に B の特徴情報を含む典型的なナビゲーション

(B) 特徴情報

B-1．典型的なサイト情報ナビゲーションを示すリンク文字列を含む

（サイトマップ，お問い合わせ，プライバシーポリシー，ヘルプ，利用規約，会社概要，採用情報，広告掲載，個人情報保護方針，会社案内，特定商取引法，免責事項，運営会社，プレスリリース，よくある質問，よくあるご質問，サイトポリシー，リンクについて，FAQ，広告掲載について，会社情報，お問合せ，初めての方へ）

(C) 検出アルゴリズム

B-1 が True なら DOM ツリーからそのコンテンツ部分を支配するタグを特定し，パンくずリストナビ領域を検出する．ただし，サイト情報ナビゲーションの場合，カテゴリナビゲーション中に出現するケースが多く存在し，2 つのナビゲーションが同じ領域に存在するような複合的なナビゲーションになりやすい．そこで，サイト情報ナビ領域前後に“句点や読点を含まない”かつ“リンクで構成されている”コンテンツである場合，サイト情報ナビゲーションを含む複合的なナビゲーションであると判断し，その領域全体を検出する．

④　ユーティリティナビゲーション

ユーティリティナビゲーションは，サイト情報ナビゲーション同様にコンテンツに対するナビゲーションではないため，どの Web サイトにおいても共通に出現しうるナビゲーションである．

そこで，サイト情報ナビゲーションと同様の方法で，典型的なユーティリティナビゲーションを示すリンク文字列を得た．しかし，クローズドテスト用の Web ページ 50 ページ中の主要部分に上記のリンク文字列を確認した所，サイトのトップページなどでは内部階層にあるコンテンツへのナビゲーションとして主要部分領域に出現する場合を確認した．また，HTML タグ上の特徴としては， JavaScript などのイベントハンドラや ASPX ファイルへのリンクなどの特徴が見られたが，広告やページ内ナビゲーションにおいても同様の特徴を持つケースがある上に，特定サイトのユーティリティナビゲーションが多数存在することから安定した検出は難しいことが分かった．

ただし，ブログページにおいては，ブログパーツなどが利用され，配置が固定される傾向にあり，主要部分との相対的な位置関係情報は重要な情報である．特に，記事の最後に出現するコメントやトラックバックを行うためのナビゲーションリンク部分の領域はブログ特有のユーティリティナビゲーションであり，検出したい情報であるため，特徴情報などを以下のようにまとめた．

(A) 検出ナビゲーション種類

A-1．ブログページ特有のナビゲーション

(B) 特徴情報

B-1．id または class 属性にブログ特有のユーティリティナビゲーションを示す文字列を含む（posted，entry_foot，postinfo）

B-2．ブログページ特有のユーティリティナビゲーションを示す文字列（コメント，トラックバック，comments，trackback）のリンク，または上記の文字列と記号や数字で構成されたリンクを含む

B-3．ブログページ特有のユーティリティナビゲーションを示す文字列を含むリンクかつリンク先アドレスが“URL#comments”または“URL#trackback”である

(C) 検出アルゴリズム

Step-1．B-1 が True なら Step4 へ

Step-2．B-2 が True なら Step4 へ

Step-3．B-3 が True なら Step4 へ

Step-4．DOM ツリーからそのコンテンツ部分を支配するタグを特定し，ブログ特有のユーティリティナビゲーション領域を検出する．

⑤　ページ内ナビゲーション

(A) 種類

A-1：ページ上部に戻るためのナビゲーション

A-2：本文へ移動するためのナビゲーション

A-3：各コンテンツへ移動するための目次型ナビゲーション

A-4：ページ末尾へのナビゲーション

(B) 特徴情報

B-1．リンク先アドレスが“#”から始まる文字列である

B-2．A-1 に頻出するリンク文字列を含む（上部，ページトップ，先頭へ等々）

B-3．A-2 に頻出するリンク文字列を含む（本文）

B-4．A-4 に頻出するリンク文字列を含む（末尾）

B-4．3 つ以上のリンクのみで構成されたタグパターンかつすべてのリンク先アドレスが“#”から始まる文字列である．

(C) 検出アルゴリズム

Step-1．B-1 かつ B-2 が True なら Step5 へ

Step-2．B-1 かつ B-3 が True なら Step5 へ

Step-3．B-1 かつ B-4 が True なら Step5 へ

Step-4．B-1 かつ B-5 が True なら Step5 へ

Step-5．DOM ツリーからそのコンテンツ部分を支配するタグを特定し，ページ内ナビ領域を検出する．Step-1 から検出した領域は A-1，Step-2 から検出した領域は A-2，Step-3 から検出した領域は A-4，Step-4 から検出した領域は A-3 のページングナビとする．

4. ナビゲーション領域検出の利用

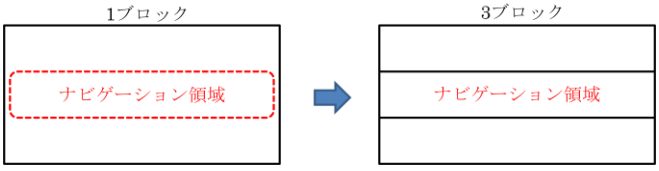
4.1 Web ページの再ブロック化

先行研究では，HTML のタグの深さに基づくコンテンツ間距離を利用し，Web ページのブロック化を行う．ここで，各ナビゲーション領域を含むブロックが存在する場合，本ナビゲーションをブロックの区切りとして利用し，ナビゲーション前後で再ブロック化を行う．

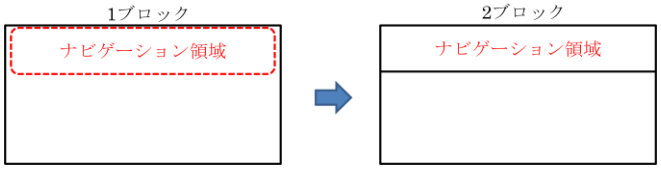
以下が，再ブロック化が必要だと考えられるナビゲーション領域の出現パターンと再ブロック化の方針である．

A）単一ブロック内にナビゲーション領域が出現

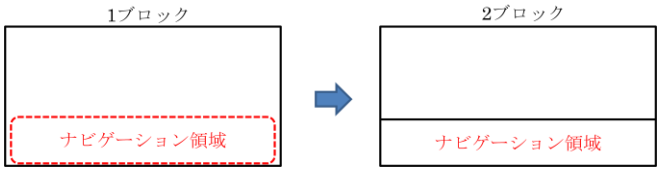
A-1：ブロック中心部にナビゲーション領域



A-2：ブロック上部にナビゲーション領域

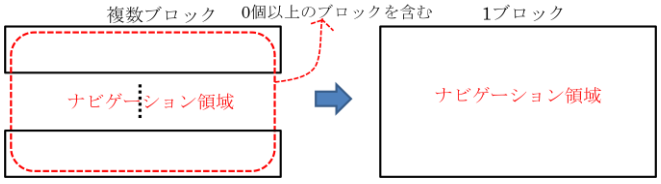


A-3：ブロック下部にナビゲーション領域

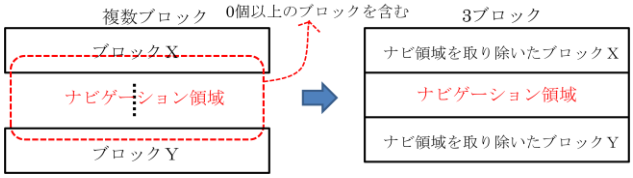


B）複数ブロック内にナビゲーション領域が出現

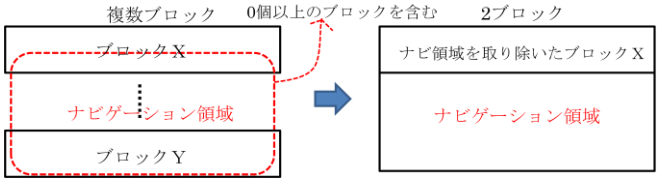
B-1：複数ブロックすべてがナビゲーション領域



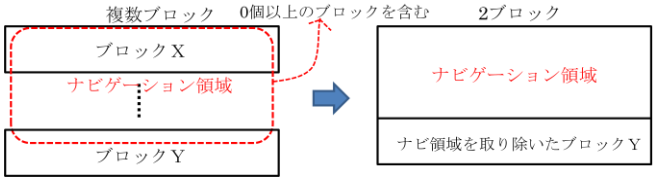
B-2：ブロック X の途中からブロック Y の途中までのナビゲーション領域



B-3：ブロック X の途中からブロック Y までのナビゲーション領域



B-4：ブロック X からブロック Y の途中までのナビゲーション領域



4.2 非主要部分特定手法

後述するクローズドテスト用の Web ページデータ 50 ページを対象とする事前調査を行い，典型的なナビゲーションの位置情報，および主要部分との位置関係を調査し，非主要部分の特定手法への利用を検討した．以下は，パンくずリストナビゲーション，ページングナビゲーション，ページ内ナビゲーションの検出情報を用いた非主要部分の特定手法である．

(1)　ナビゲーション領域検出情報を利用した処理

ページ中に検出したナビゲーションごとに①~⑤の処理を行う．各ナビによるブロック同定処理が終了したら(2)へ．

①　パンくずリスト検出処理

I.　(1)でパンくずリストナビゲーションを検出した場合，検出したナビ数により処理を分岐する．

パンくずリストナビ検出数が 1 つの場合，Ⅱへ
パンくずリストナビ検出数が 2 つの場合，Ⅲへ
パンくずリストナビ検出数が 3 以上の場合，Ⅳへ

II.　以下の処理を行う．

パンくずリストナビゲーションはヘッダ領域の直後に配置される傾向にあるため，パンくずリストナビゲーション以前に出現するブロックは非主要部分と同定する．また，パンくずリストナビゲーション領域以降に出現する一番強い<h>タグ見出しは主要コンテンツに対する見出しである傾向にあるため，その見出しを含むブロックは主要部分と同定する．ただしブロックが見出し単体の場合は，その次のブロックも主要部分と同定する．

(1)に戻る．

III.　パンくずリストナビゲーションの種類により，処理を行う．

同一のパンくずリストナビゲーション同士の場合,パンくずリストナビゲーションはヘッダ領域の直後とフッタ領域の直前に配置される傾向にあるため，パンくずリストナビゲーションに挟まれた領域以外のブロックは非主要部分と同定する．また，パンくずリストナビゲーション間の領域に出現する一番強い<h>タグ見出しは主要コンテンツに対する見出しである傾向にあるため，その見出しを含むブロックは主要部分と同定する．ただしブロックが見出し単体の場合は，その次のブロックも主要部分と同定する．

異なるパンくずリストナビゲーション同士の場合,どちらかのナビが誤検出である可能性が高いため，ここでは同定処理は行わない．

(1)に戻る．

IV.　パンくずリストナビ種類により，処理を行う．

同一のパンくずリストナビが 2 つ＋異種のパンくずリストナビの場合，同一のパンくずリストナビを優先し，同種のパンくずリストナビはヘッダ領域の直後とフッタ領域の直前に配置される傾向にあるため，パンくずリストナビに挟まれた領域以外のブロックは非主要部分と同定できる．

上記以外の場合,同種のパンくずリストナビが 3 つ出現するケースなどが考えられるが，そういったページが出現する頻度はかなり低く，パンくずリストナビの誤検出の可能性も高いため,ここでは処理を行わない．

(1)に戻る．

②　ページングナビゲーション検出処理

I.　(1)でページングナビゲーションを検出した場合，ナビゲーションの数により処理を分岐する．

本ナビゲーションの検出数が 1 つの場合，Ⅱへ
本ナビゲーションの検出数が 2 つの場合，Ⅲ
本ナビゲーションの検出数が 3 以上の場合，Ⅳへ

II.　ページングナビゲーションの種類により，処理を行う．

主要部分の直後に配置されるため,本ナビゲーションの直前ブロックは主要部分，本ナビゲーション以降のブロックは非主要部分だと同定する．直前ブロックが他のナビゲーションと判定されている場合，繰り上げてさらに前のブロックを主要部分と同定する．

(1)に戻る．

III.　ページングナビゲーションの種類により，処理を行う．

同種のページングナビゲーション同士の場合,主要部分の直前及び直後に配置される傾向にあるため，ページングナビゲーション間のブロックは主要部分だと同定し，それ以外のブロックを非主要部分と同定する．

異なるページングナビゲーション同士の場合,どちらかのナビゲーションが誤検出である可能性が高いため,ここでは同定処理は行わない．

(1)に戻る．

IV.　ページングナビゲーションの種類により，処理を行う．

同種のページングナビゲーションが 2 つ＋異種のページングナビゲーションの場合，同種のページングナビゲーションを優先し，同種のページングナビゲーション間のブロックは主要部分だと同定し，それ以外のブロックを非主要部分と同定する．

上記以外の場合,同種のページングナビゲーションが 3 つ出現するケースなどが考えられるが，そういったページが出現する頻度はかなり低く，ページングナビゲーションを誤検出している可能性も高いため，ここでは処理を行わない．

(1)に戻る．

③　サイト情報ナビゲーション検出処理

I.　(1)でサイト情報ナビを検出した場合，著作権表示や id,class 属性に header, footer がセットされているかなどの情報をもとに出現位置パターンを特定し，処理を分岐する．

出現位置がヘッダ領域の場合，Ⅱへ
出現位置がサイド領域の場合，Ⅲへ
出現位置がフッタ領域の場合，Ⅳへ

II.　以下の処理を行う．

ヘッダ領域に配置されるため,本ナビよりも上部のブロックを非主要部分と同定する．

(1)に戻る.

III. 以下の処理を行う.

サイド領域に配置されるが,関係左右どちらに配置されるかを特定するのが困難であり，主要部分との相対的な位置関係は固定されないため，この位置情報からの同定は行わない.

(1)に戻る.

IV. 以下の処理を行う.

フッタ領域に配置されるため,本ナビよりも下部のブロックを非主要部分と同定する.

(1)に戻る.

④ ユーティリティナビゲーション検出処理

I. (1)でブログ特有のユーティリティナビゲーションを検出した場合，ナビゲーションの数により処理を分岐する.

本ナビゲーションの検出数が 1 つの場合，Ⅱへ

本ナビゲーションの検出数が 2 つ以上の場合，Ⅲへ

II. 以下の処理を行う.

ブログ記事（主要部分領域）の最後に出現するため,このナビ以降のブロックを非主要部分と同定する．また，本ナビより前に出現する<h>タグは，記事タイトルである傾向にあるため，その<h>タグを含むブロックから本ナビ間のブロックを主要部分と同定する.

(1)に戻る.

III. 以下の処理を行う.

各ブログ記事(主要部分領域)の最後に出現するため,最後に出現した本ナビ以降のブロックを非主要部分と同定する．また，各ブログ記事に存在する本ナビより前に出現する<h>タグは各記事タイトルである傾向にあるため，<h>タグを含むブロックからナビ間のブロックをそれぞれ主要部分と同定する.

(1)に戻る.

⑤ ページ内ナビゲーション

I. (1)でページ内ナビを検出した場合,検出したナビ種類により処理を分岐する.

ページ上部に戻るためのナビゲーションの場合,Ⅱへ

本文へ移動するためのナビゲーションの場合，Ⅲへ

各コンテンツへ移動するための目次型ナビゲーションの場合，Ⅳへ

ページ末尾へのナビゲーションの場合，Ⅴへ

II. 以下の処理を行う.

移動先はヘッダ領域,またはその上部がヘッダ領域であるので，移動先のブロック以前のブロックを非主要部分と同定する．また，移動先ブロックとブロック終端までにコンテンツが存在しない場合，移動先ブロックも非主要部分と同定する.

また，本ナビが 1 つの場合，主要部分領域終了後のページ下部に存在するため，本ナビゲーション以降のブロックは非主要部分と同定する.

本ナビが複数の場合,複数の主要部分ごとに配置されていることから，最後に出現したナビの位置が主要部分領域終了であると判断し，最後に出現した本ナビゲーション以降のブロックは非主要部分と同定する.

(1)に戻る.

III. 以下の処理を行う.

移動先が主要部分領域の始まりであるので,移動先のブロックを主要部分と同定する．ただし，移動先とそのブロックの終端にコンテンツが存在しない場合，その次のブロックを主要部分と同定する.

(1)に戻る.

IV. 以下の処理を行う.

各移動先が主要部分領域の始まりであるので,移動先のブロックを主要部分と同定する．ただし，移動先とそのブロックの終端にコンテンツが存在しない場合，その次のブロックを主要部分と同定する.

(1)に戻る.

V. 以下の処理を行う.

各移動先がページの終端またはフッタ領域の始まりであるので，移動先のブロックを非主要部分と同定する．ただし，移動先とそのブロックの終端にコンテンツが存在しない場合，その前のブロックを主要部分と同定する.

(1)に戻る.

(2) 未特定ブロックに対する処理

(1)で特定されていないブロックについては，クローズドテスト用データセットを用いて 2.2 節の先行研究手法と同様に C4.5 による主要･非主要部分の決定木学習を行い，出力された決定木にかけて，主要か非主要かを判定する.

5. 評価実験

5.1 データセットの収集・作成

ウェブページ 10 万ページを取得し形態素解析を行い, TF/IDF で上位となった語から自然な組み合わせを選ぶという手法によって 2 語のキーワード対を 118 組用意する．このキーワード対について検索エンジン google を用いて検索を行い，検索結果のうち上位 10 ページずつ合計 1180 ページを用意する.その中からサイズ順データの中間層から 200 ページを抽出し，0 バイトのページやドメインが同じページを削除し，減少した分は追加する．また，キーワードの偏りをなくすために，キーワード対ごとに最大で 3 ページとなるように追加，削除を行う.

上記の方法で収集した 200 ページに対して，大学生 2 名による各ナビゲーションおよび主要部分のタグ付けを行い, 50 ページをクローズドテスト用データ, 150

ページをオープンテスト用データとした.

5.2 非主要部分特定性能の評価

5.1 節の手順で作成したデータセットを対象に，非主要部分特定性能のクローズドテスト，及びオープンテストを行った．実験結果の表中，Hit は非主要部分を正しく判定できたもの，FA(False Alarm)は非主要部分でないものを誤って判定したもの，Miss は非主要部分を取りこぼしたもの,精度は Hit/(Hit+FA),再現率 は Hit/(Hit+Miss), F 値 は精度と再現率の調和平均であり，総合的な性能指標である.

表 5.3.1　先行研究手法での評価結果

	Hit	FA	Miss	精度(%)	再現率(%)	F値(%)
ClosedTest	365	156	18	70.1	95.3	80.8
OpenTest	977	547	140	64.1	87.5	74.0

表 5.3.2　提案手法での評価結果

	Hit	FA	Miss	精度(%)	再現率(%)	F値(%)
ClosedTest	503	133	29	79.1	94.5	86.1
OpenTest	1419	510	100	73.6	93.4	82.3

先行研究[1]での手法と提案手法による非主要部分特定性能を上述のデータセットを用いて比較した．クローズドテストにおいては，本手法の F 値が先行研究の手法より 5.4 ポイント向上している．また，オープンテストにおいては，本手法の F 値は先行研究の手法より 8.3 ポイント向上している.

5.3 考察

検出ナビゲーションの組み合わせ別の評価を行ったところ，クローズドテスト，オープンテストでの F 値が最も向上したのはブログ特有のユーティリティナビゲーション(以下,ユーティリティナビゲーション）以外を検出した場合であった．また，ユーティリティナビゲーションは，単体のみ検出した場合でのオープンテストの F 値が大きく低下したため，ナビゲーションによる特定と決定木による特定手法別での性能内訳を確認した．その結果，ユーティリティナビ検出を利用した特定性能は向上しており，F 値が低下しているのは，決定木によるブロック特定性能の問題であることが分かった．つまり，ユーティリティナビ検出を利用して Web ページの再ブロック化したものを上手く決定木学習できなかったことが原因である．特に，ブログ記事内に画像コンテンツなどが存在する場合にそこでブロックの区切りとなってしまうケースがあり，さらにユーティリティナビによる再ブロック化を行ったことで主要部分が細分化された状態で学習させてしまったことが悪影響を及ぼしたと考えられる.

上記のような問題に対処するためには，ページタイプを考慮したナビゲーション領域の特定処理を行うことが有効だと考えられる．特にブログページでは，主要部分・非主要部分がはっきりしており，位置やレイアウトもブログサービスによってある程度固定されるため，ブログページ処理などの強力なヒューリスティクスな手法を利用することも有効であると考えられる.また，ナビゲーション領域はサイトの種類によって出現する頻度が大きく変わるため，ページタイプごとのナビゲーション領域処理を変更することで性能向上の可能性もある．例えば，ページングナビゲーションは EC サイトなどの商品検索結果の表示やニュースサイトの記事の分割のために利用されることが多いため，そういったページではページングナビゲーションの処理を優先させるといったことが考えられる． 今後，ページタイプごとのナビゲーション領域の出現傾向や特徴について，検討していく必要がある.

6. おわりに

本稿では，まず， Web ページを対象とした主要部分・非主要部分特定手法について議論し，Web ページ群を対象とした手法と単一の Web ページを対象とした手法について述べた．特に，単一の Web ページを対象に機械学習を用いた手法として先行研究[1]を挙げ，既存手法の中で着目されていないナビゲーション領域というものに注目した.

本研究では，上記のナビゲーション領域を定義，検出する手法について検討を行い，検出性能について評価を行った．また，ナビゲーション領域検出情報を用いた Web ページの再ブロック化, 非主要部分特定手法を提案し，先行研究の手法と組み合わせることで，先行研究の手法よりも F 値を 8.3 ポイント向上させるという成果を得られた．また検出ナビゲーション別の評価実験などにより，ナビゲーション領域ごとの有効性や改善の余地について確認でき，さらなる性能向上のための知見を得た.

参　考　文　献

[1] 齋藤佳枝，”Web 中の非主要部分特定手法”，静岡大学情報学部卒業論文 2008.

[2] 服部元，松本一則，菅谷史昭，”タグの深さを利用したコンテンツ間距離に基づく Web ページの自動分割方式”，日本データベース学会 letters，Vol.4，No.1，pp.149-152，2005.

[3] Shian-Hua Lin and Jan-Ming Ho,”Discovering Informative Content Blocks from Web Documents”, In Proceedings of ACM SIGKDD 2002 , pp.588-593,2002.

[4] 吉田光男，山本幹雄，”教師情報を必要としないニュースページ群からのコンテンツ自動抽出”，日本データベース学会論文誌，Vol.8，No.1，pp.29-34，2009.

[5] 鶴田雅信, 増山繁, ”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，pp.197-200，2008.

[6] 鶴田雅信, 増山繁, “レイアウト情報を用いた Web ページの主要な DOM ノードの抽出法”，人工知能学会論文誌，Vol.25，No.6， pp.742-756，2010.

[7] 新谷剛司ほか，” すべての人に知っておいてほしい web デザインの基本原則”，MdN，2011.