| | |
|---|---|
| Title | 視覚障害者用音声ブラウザのためのウェブページ構造解析 |
| Author(s) | 加藤，邦彦 |
| Citation | |
| Issue Date | 2006-03 |
| Type | Thesis or Dissertation |
| Text version | author |
| URL | http://hdl.handle.net/10119/1965 |
| Rights | |
| Description | Supervisor:白井　清昭，情報科学研究科，修士 |

修 士 論 文

# 視覚障害者用音声ブラウザのための ウェブページ構造解析

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

加藤 邦彦

2006 年 3 月

修 士 論 文

# 視覚障害者用音声ブラウザのための ウェブページ構造解析

指導教官　白井 清昭 助教授

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

410030 加藤 邦彦

提出年月: 2006 年 2 月

**概 要**

近年の WWW の発達により, 視覚障害を持つ人であっても計算機を介してさまざまな情報を得ることができるようになっている. 視覚障害者はページの文字情報を音声で読み上げるツールを用いてウェブブラウジングを行っているが, 既存のツールには知りたい情報が書かれている場所に辿りつくまでに手間や時間がかかるといった問題点が存在する. したがって本稿では,(1) ウェブページの構造解析を行い, その構造に基づいてページのセグメンテーションを行い, 不要な情報を読み飛ばす,(2) ユーザがリンクを辿ったとき, リンク先ページの中からユーザが知りたい情報が書かれている位置を推定し, その場所から読み上げを開始する, という二つのシステムの提案を行うことによって, より適切に視覚障害者のウェブブラウジングを支援する新たなツールの開発を行っている. そのために DOM を用いたウェブページセグメンテーション手法やリンクが指すページ内位置の同定手法を考案し, それらの手法の有効性を確認するため, 実験を行った.

# 目 次

# 図 目 次

# 表 目 次

# 第1章　はじめに

近年のWWWの発達により, インターネットを利用し様々なことを行うことができるようになっている. 視覚障害を持つ人であってもウェブを利用することによって, 自宅に居ながらにして様々な情報を得ることができるようになった. 例えば, ウェブショップを通して買い物をしたり, 掲示板を用いて他の人と情報交換を行ったりしている. それらの人々は画面の情報を音声化するソフト (スクリーンリーダー, 音声ブラウザ) を用いてネットブラウジングを行っている. しかし, それらのソフトはHTMLソースのタグ情報を基にテキストを順番に読み上げているだけであるため, ユーザの目的の箇所まで到達するのに時間と手間がかかる. また, 表示されている文字情報を単純に音声化するだけではページ全体の構造を把握することは難しい. よってGUI(グラフィカルユーザインターフェイス) を使用することができない視覚障害者のネットブラウジングの効率的な支援ができるシステムの開発が望まれている.

## 1.1　視覚障害者のためのツール

現在の視覚障害者支援を目的としたツールには, スクリーンリーダーや音声ブラウザなどがある. スクリーンリーダーとはウェブブラウザに限らず, 計算機の画面上に表示されている情報を読み上げるツールで, 汎用性が高く様々なアプリケーションに対応している. ウェブブラウジングを行う際には, 一般に使用されているInternet Explore等のブラウザを利用し, そこで表示されている情報を音声化する. 従って画面上に表示されている情報は, 一般の人が画面を通して得る情報と全く同じである. しかし視覚障害者はマウスを用いるGUIを利用することは困難なため, それらのブラウザに対する操作はキーボードを用いて行われる. その際の操作方法は各々ウェブブラウザによって異なるが, 大抵の場合テキスト間の移動はタブキーやカーソルキーなどを用いて, クリック操作はエンターキー用いて行う. 代表的なものとして, 高知システム開発社の「PC-Taker」や日本IBM社の「JAWS」などがある.

一方音声ブラウザとは, ウェブページを音声化することに特化したツールである. スクリーンリーダーとの違いは, スクリーンリーダーはウェブブラウジングを行うためのブラウザを別途必要とするのに対し, 音声ブラウザはそれだけでウェブブラウジングを行うことができる. またテーブルの読み上げ機能やリンクを読み上げる際には声質が変わるといった,HTMLタグを利用したウェブページ特有の機能が備わっている. ユーザインターフェイスに関しては, スクリーンリーダーと同様にキーボードを用いて行われている. 代表的なも

のとしては, 日本 IBM 社の「ホームページリーダー」が挙げられる.

しかし, これらのツールは冒頭でも述べたように, 単純にソースに書かれているテキストを読み上げていくだけであるため, ページを全て読むには, また目的の情報まで辿り着くには多大な時間と労力がかかるという問題点があり, その問題点はしばしば視覚障害者のウェブ利用の妨げとなっている.

## 1.2 提案システム

本研究では実際に視覚障害者の立場でウェブブラウジングを行っている方からの意見を頂き, 新しい支援システムの構築を目指した. 本研究で提案する視覚障害者支援システムの概要は次のようなものである.

1. ウェブページの構造を自動解析し, セグメントの検出を行い, それに基づいて不必要な箇所の読み飛ばしをナビゲートする.

2. ユーザの目的位置が特定できる場合には, ページの読み上げをその箇所から開始する.

これらの機能を搭載したウェブブラウザを開発することによって, 視覚障害者のウェブブラウジングをより適切に支援し, 音声情報のみでもより効率よく支援することができると考えられる.

本稿ではそれらの支援システムの提案を行い, さらにそのシステムの中で使用されるウェブページセグメンテーション手法やリンクの参照位置同定の手法を紹介する. 章の構成は次のとおりである.2 章では本研究の関連研究について,3 章では本研究で提案したシステムの概要について述べる.4 章では本研究で用いたウェブページセグメンテーション手法を,5 章では目的位置の同定を行う手法について紹介する. 最後に 6 章でまとめを述べる.

# 第2章　関連研究

## 2.1　音声ブラウザ

第1章でも述べたように視覚障害者がウェブブラウジングを行う際には, ウェブページの文字情報を音声情報に変換する音声ブラウザと呼ばれるツールが用いられる. 本節では音声ブラウザの中の代表的なものについて紹介する.

代表的な音声ブラウザとして, 日本IBM社の「ホームページ・リーダー」が挙げられる. このソフトには, テーブル読み上げ機能や, テキストとリンクで読み上げの声を変化させる機能が備わっている. ルビ要素やアクセスキー機能への対応などを進め, ウェブアクセシビリティの標準や規格に対応したページを, より良く操作できるようになっている. 近年多くの場面で使用されるAdobe PDF文書やマクロメディアのFlashコンテンツにも対応し, これらを簡単な操作で読み上げることができる. また見出し, テーブル, データ入力域を追いかけるモードを提供しており, ユーザが高速な操作ができるよう配慮されている. これらの操作はすべてキーボード上で行うことができる. さらにテキストのみの画面ではなく, レンダリングされたページも表示するため, ページのアクセシビリティ検査に利用することも可能である. 本研究では本ソフトを実際に使用し, さらに視覚障害者の方にも使用して頂き, その中で浮かび上がった問題点を解決するための手法を提案している.

他の音声ブラウザとしてアメディア社の「ボイスサーフィン」がある. 前述したホームページリーダーと同様にページの読み上げをキーボード操作で行うことができ, 見出しへのジャンプや声の変化といった機能が備わっている. このソフトの特色として,「記事本文アクセス」機能が挙げられる. この機能は, 新聞社のサイトで記事の本文から読み上げを行うための機能として開発され, 記事のタイトル部分のリンクを選択すると, 該当のページを読み込み, 記事本文の先頭を探して読み上げを始める機能である. 本研究においても同様のシステムの開発を行っている.

## 2.2　視覚障害者支援に関する研究

WWWに関する視覚障害者支援を目的とした研究としては前田らの研究が挙げられる[1]. この論文ではユーザの視覚特性に適応してウェブページの文字サイズや背景の色などを変換するシステムが示されていた. ユーザの視力, 色盲といった視覚特性に関する情報を蓄積し, その情報に基づいてウェブページの文字サイズや色, 背景色を自動変換するシステムである. しかしこれらは弱視者を対象としたシステムで全盲のユーザを対象としたもの

となってはいない.

また音声ブラウザに関する研究として浅川らの研究が挙げられる [2]. 浅川らはホームページリーダーと呼ばれる音声ブラウザを開発した. また, そのシステムにおいて, テンキーを用いてユーザとやり取りを行うことや, 読み上げを行う際に HTML タグを音声にどのように変換させるかについて紹介している. 現在まで, このホームページリーダーの開発は続けられており, このツールが視覚障害者のウェブブラウジングに対して有益な支援を行うことができることが実証されている. しかし近年の著しいウェブ技術の進歩により,HTML の多様化, ユーザの欲求の変化といった問題が発生しており, 新たな技術の導入が必要となっているのも事実である.

## 2.3 ウェブページ構造化に関する研究

ウェブページの構造認識を目的とした研究として, 南野らの研究がある [3]. 南野らは, タグ情報の繰り返しから適切な構造の境目を判別し, それを基に構造認識を行う手法を提案している. 最下層のタグの繰り返しを検出し構造化したならば, その構造をグループ化しさらに上の階層で繰り返しを検出する. これをボトムアップに繰り返すことで構造化を行う. 繰り返しの判定には DP マッチングを利用しており, 完全一致で構造化できなかったものに対して DP マッチを行って再度構造化を試みている. この手法では DP マッチを利用したことにより, 非常に計算時間がかかる場合が存在したことが報告されていた. ウェブブラウジングを行っている際に余りに多くの計算待ちの時間がかかるなら, ユーザに対して適切な支援を行っているとは言えないため, この手法を視覚障害者支援にそのまま適用することは困難である. また個人サイトのウェブページなど繰り返し構造が明示的でないものも多く, 繰り返し構造のみでページの構造化を行うのは困難といえる. しかしながら繰り返し構造はページの構造を検出する上で重要な特徴となるため, 本研究でもその他の手法に加えて繰り返し構造に基づいた手法をセグメント検出に適用している.

Yu らは,Dcument Object Model(DOM) の情報と視覚情報を組み合わせてページの意味構造を認識するシステムを提案している [4]. 彼らは DOM 情報に加えて視覚情報を利用してウェブページをセグメンテーションする VIPS アルゴリズムを導入している.VIPS アルゴリズムではタグ, 色, サイズといった特徴に基づいてセグメントを抽出し, セグメントが検出されなかった箇所では, 距離, タグ, フォント, 色のパターンにより水平, 垂直両方向のセパレータを検出している. そして, それらを組み合わせることによってウェブページを構造化している. 彼らは構築した構造を疑似関連フィードバック法に利用し, ウェブ検索性能が向上したことを実証した. 本研究ではこれらの構造化の手法を音声ブラウザのために利用するため, 彼らの手法に新たな手法を組み合わせ, 視覚障害者支援システムへの適用を行っている. 彼らの手法の目的は検索性能の向上であり, 文章を多く含んだウェブページがページセグメンテーションの対象となっていたが, 本研究では様々な種類のウェブページにおいてセグメンテーションできる手法が必要であるため, さらにシステムの改良が求められる.

松本らは,DOM 構造や繰り返し構造を用いて構造化できないウェブページの構造化に取り組んでいる [5] . 彼らはテキストセグメント階層構造が日本語の係り受け構造と類似していることに着目し,3 値分類による構築アルゴリズムと 2 段階構築アルゴリズムの二つのアルゴリズムを提案している. そのアルゴリズムにおいてテキスト間の係り関係推定を行う際に DOM のパスやインデント情報, 言語的情報を利用している. ページのセグメンテーションに関しては, 基本単位はブロックレベル要素とされており, 装飾などに用いられるインラインレベル要素は削除している. また, この研究の目的は構造構築に焦点を当てたものであるため, セグメントの抽出誤りは人手で修正している. 本研究ではウェブページのセグメントの検出に焦点を当てており, ブロックレベル要素だけではなく, 文字装飾などのインラインレベル要素も利用している.

Yang らは, ウェブページのコンテンツの視覚的類似点に基いてページの意味構造を分析する手法を提案している [6]. 彼らは人間がレンダリングされたウェブページを見て認識する意味構造を自動獲得することを目指しており, 視覚的類似点において頻出するパターンを検出するために, ヒューリスティックなパターン検出アルゴリズムを適用している. そのアルゴリズムによってパターンを検出した後, そのパターンをサフィックスツリーで表し, そのツリーを用いてクラスタリングを行っている. 彼らは提案手法を用いてウェブページの視覚情報に基づいた意味構造を構築でき, ウェブページコンテンツの省略などに利用できると述べている.

Yoshida らは, ウェブページの HTML テーブルからシソーラスを抽出する手法を提案している [7]. 彼らは様々な形の HTML テーブルから属性と値となる部分を識別, 抽出し, ひとつのテーブルにクラスタリングするシステムを開発した. そのシステムは EM アルゴリズムを用いたもので, 教師なし学習によってテーブルの構造を認識し, その構造に基づいて属性と値の識別を行っている. 本研究では HTML ソース中のテーブルをひとつのセグメント要素として利用しているが, より効率的な視覚障害者支援を行っていくうえで, 彼らの研究で行われているようなテーブル中の属性と値の識別を行っていく必要がある.

本研究では, これらの手法を取り入れつつ, 新たな手法の開発を目指してきた. 適切にウェブページの構造解析を行うことにより, 視覚障害者がウェブブラウジングを行う際に発生する手間を省略し, 音声のみの情報でも構造の把握を容易にするといった, より適切な支援ができると期待できる.

# 第3章　提案システム

この章では, 本研究で作成する支援ツールの概要を述べる.

## 3.1　ウェブページのセグメントに基づくナビゲーション

既存のツールの最大の問題点は, 音声のみを用いてユーザに情報を伝えるため, 手間と時間がかかってしまうことである. それはシステムが読み上げを行う際に,HTML ファイルのソースに書かれているテキスト等をソースの出現順序に従って単純に読み上げを行っていることに起因している. この点はウェブページの内容理解に関しても問題を発生させる. つまりウェブページ内にあるテキストを単純に読み上げるだけでは, 視覚障害者がページの内容を理解することは困難になり, 何度も同じページを開かなければならなくなる. よって本研究では, ウェブページを構造情報に基づいていくつかのセグメントを検出し, そのセグメントを利用して読み上げの際に不必要なテキストを読み飛ばすことによって, 視覚障害者のウェブブラウジングに必要な手間を省略する.

例えば既存のツールを使用した場合, 図 3.1 のウェブページ中央部分にある「電話帳」にたどり着くには上部から順番にたどって行かなければならない. その場合, 音声しか認識できない視覚障害者にとっては与えられる情報量が多くなり, ページがどのような構造になっているかを把握することも困難である.

そこで本研究で開発するシステムは, 図 3.2 のようにウェブページのセグメントを検出する. 図 3.2 中の枠線が検出されたセグメントである. このセグメントに基づいて読み上げ支援を行う. 本システムの手順は次のとおりである.

1. ページのセグメントの検出を行う
2. 読み上げの際セグメントの箇所では, まずユーザに見出しなどのセグメントの内容を示す情報を提示し, その内容を読むか, スキップするかを選択させる
3. スキップが選択されたならば, そのセグメントの中に含まれるテキストを読み飛ばし, 次のテキストを読み上げる

以上の手順によって, ユーザがページの閲覧を行う際にユーザに対して与えられる情報が省略され, 必要な手間を減らすことができると考えられる.

先ほどの例であれば, 図 3.2 を見ると「電話帳」というコンテンツは「調べる」というセグメントの中に存在することになる. これらのセグメントに従ってページ読み上げを行う


図 3.1: ウェブページの読み上げの例 (システム未使用)

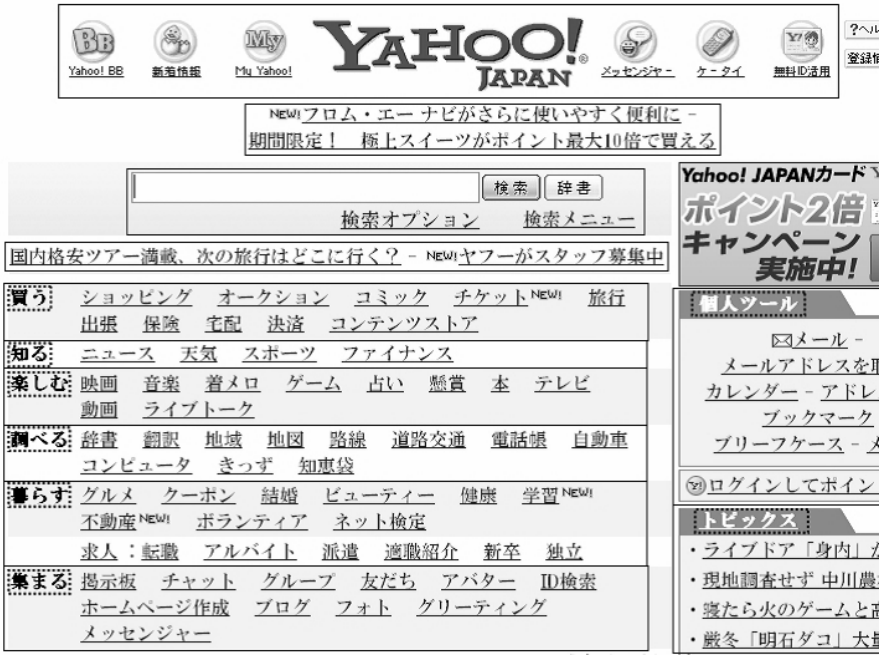


図 3.2: ウェブページの読み上げの例 (システム使用)

場合, システムは「買う」という見出しを持ったセグメントがあることをユーザに知らせ, ユーザはそのセグメントを読み上げないと選択することによって「ショッピング」,「オークション」,「コミック」といったコンテンツを読み飛ばすことができる. 同じようにして「知る」,「楽しむ」といったセグメントを読み飛ばし, 素早く「調べる」セグメントの中にある「電話帳」まで辿り付くことができる.

これらのシステムを実装するためには, まずウェブページのセグメントを検出する必要がある. そこで本研究では, システムによって検出すべきセグメントを次のように定義した. これらの定義は, ページの読み上げを行う際に有効であるという観点から作成されている.

- 意味のあるまとまり (共通の属性を持つ)
- レイアウト上の固まり
- 要素が 2 つ以上
  - 短すぎるセグメントを作成したとしても, ページの読み飛ばしに有効ではない
- セグメントの先頭に見出しがあることが望ましい
  - ナビゲートする際に見出しを活用できるため
- 各セグメントは入れ子になっていても良い
- ページ全体をカバーしなくても良い
  - ページの構造構築ではなく, ページコンテンツの省略を目的としているため

これらのセグメントを検出するための詳しい手法に関しては, 第 4 章で述べる.

## 3.2 ユーザの目的位置から読み上げを開始するシステム

ニュースページの記事を示したリンクをユーザがクリックした場合, ユーザは新たに開かれるページの中の記事について関心があると考えることができる. その際, 新たに開かれたページをもう一度最上部から読み上げていくことはユーザにとって余分な手間をかけさせることになる. 従ってユーザが知りたいと思われる箇所を特定できた場合には, ユーザに通知し, 望むならばその箇所から読み上げを開始することによって, ウェブページの閲覧をより適切に支援することができる.

システムの手順は, 以下のとおりである.

1. 親ページでクリックしたアンカー文字列と類似する文字列を, リンク先である子ページの中から探索する

2. 類似する文字列が検出された場合には, ユーザに類似した文字列を検出したことを通知する

3. ユーザが望めばその箇所から読み上げを開始する

例として, 図 3.3 と図 3.4 を挙げる. 図 3.3 のページの中でひとつの記事をクリックした場合, 既存のツールでは新たに開かれた図 3.4 ページの読み上げは最上部から開始される. 従ってユーザは HMTL ソースの前部にある「新聞購読」,「電子有料版」といった上部のコンテンツや「社説など」,「朝刊から」などの左部のコンテンツを経て, 記事まで辿り着かなければならない.

しかし本システムを使用するならば, ユーザが親ページにおいて記事をクリックし新たなページが開かれた時点で, システムはアンカー文字列を利用してページ内のリンクが指す箇所を探索する. その探索でページ内位置が同定されたならば, ユーザに対して通知され, システムはその箇所から読み上げを開始するのでユーザは即座に自分が必要とする情報を得ることができる.

なお, 同様のシステムが「ボイス・サーフィン」の中で実装されている. 本研究ではシステムを実際に作成し本システムの有効性を検証すると共に, 既存のソフトの中では検出できなかった事例を解析し, それらを検出することのできるシステムの構築を目指した.

リンクが指すページ内参照位置の同定手法は第 5 章で詳述する.


図 3.3: リンクの参照位置 (親ページ)


図 3.4: リンクの参照位置 (子ページ)

# 第4章　ウェブページセグメンテーション

3.1 節で述べたウェブページの読み飛ばしを行うシステムは, ウェブページのセグメントに基づいてナビゲートを行うため, ページのセグメント検出が重要な要素となる. 従ってページの構造に基づいた適切なセグメンテーションを行う必要がある.

本章では, 本研究において使用したウェブページセグメンテーションの手法について紹介し, 実際のウェブページを用いて行った実験の結果を示す.

## 4.1　提案手法

本研究で作成したシステムの大まかな流れは次のようになっている.

1. 前処理 (XML への変換, タグの除去)
2. DOM 構造を用いて大きなセグメントを取り出す
3. 大き過ぎるセグメントの分割を行う
4. 見出しと内容が分かれているセグメントのマージを行う

### 4.1.1　前処理

この説では, 本研究で作成したシステムの実行のために必要な前処理について述べる.

#### XML への変換

本システムは perl の DOM モジュールを用いて実装した.perl の DOM モジュールを使用する場合,HTML 文書を XML 文書に変換する必要がある. そのため本研究では Tidy[8] と呼ばれる HTML 文書自動訂正ツールを使用した.Tidy とは HTML ソースが w3c の規定に準拠しているかを検査するもので, これを用いると, ソース中の HTML タグの整合性を検査し, 整合性が取れていない箇所のタグを整える.Tidy は以下の働きをすることにより,HTML 文書の訂正を行う.

- 開始タグと終了タグが必要な HTML タグにおいて, 終了タグがないことを検知すると自動的に終了タグを適切な位置に挿入する

- 終了タグの出現順序が開始タグと対応していない場合には, 対応する位置に終了タグを修正する

また Tidy には asxml オプションを使用することによって HTML 文書を XML 文書に自動変換する機能も備わっている. 本研究ではそのオプションを利用し,HTML 文書から XML 文書への変換を行った.

#### タグの除去

Tidy による文書訂正を行う際にある特定のタグがエラーを引き起こすことが判明したため,Tidy による処理を行う前に以下のタグを HTML ソースより除去した.

- `<table>`タグの中にあるフォーム (`<form> </form>`)
- スクリプト (`<script> </script>`,`<scr> </scr>`)
- スタイルシートを使用する部分 (`<style> </style>`)

### 4.1.2 Document Object Model(DOM) を用いたセグメンテーション

本研究で提案する手法は DOM を利用したものである. 前処理で Tidy により HTML を XML に変換した後,XML パーザの中に含まれている DOM のモジュールを用いることにより,HTML 文書を木構造で表現する. その木構造より下記のタグを用いてセグメントを検出する.

- タグで囲まれている箇所をセグメントとみなす ($T_b$)
  - table, ol, dl, ul, p
- それ自体をセグメントの境界とみなす ($T_l$)
  - h1, h2, h3, h4, h5, h6, hr

セグメンテーションプログラムのアルゴリズムは下記のとおりである. アルゴリズムはセグメンテーションに利用するタグの探索を行う検出ステップと, 実際に HTML ソースにセグメント情報を挿入する挿入ステップに分かれている.

- 検出ステップ

  **for all** $tag_i \leftarrow$DOM ツリー **do**
    **if** $tag_i = T_b$ **then**
      $nest \leftarrow check\_nest(tag_i)$
      **if** $nest =$ YES **then**

$tag_i$ を無視する
**else if** $nest$ = NO **then**
$tag_i$ にセグメントマークを付ける
$ST(tag_i) \leftarrow$ Block
**end if**
**else if** $tag_i = T_l$ **then**
$tag_i$ にセグメントマークを付ける
$ST(tag_i) \leftarrow$ Line
**end if**
**end for**

- 挿入ステップ

**for all** $tag_i \leftarrow$DOM ツリー **do**
**if** $tag_i$ = セグメントマーク **then**
**if** $state$ eq open **then**
$tag_i$ の直前に終了キューを挿入
**end if**
**if** $ST(tag_i)$ = Block **then**
$tag_i$ の前後に開始, 終了キューを挿入
$state \leftarrow$ close
**else if** $ST(tag_i)$ = Line **then**
$tag_i$ の直前に開始キューを挿入
$state \leftarrow$ open
**end if**
**end if**
**end for**

アルゴリズム中の変数を以下に説明する.

- $tag_i$:DOM ツリーにある各タグ
- $nest$:入れ子状態を示すもの (YES or NO)
- $ST(tag_i)$:$tag_i$ のセグメントタイプ (Block or Line)
- $state$:現在のセグメントの状態を表す (初期値:close)
  - 終了キューを挿入していないセグメントがある場合,open
  - 終了キューを挿入していないセグメントがない場合,close

検出ステップで, システムはセグメンテーションに利用できるタグを検出する.DOM ツリーを上から探索し, そこに存在する各 HTML タグが $tag_i$ 変数に代入され,$tag_i \in \{T_b, T_l\}$ となる $tag_i$ を検出する. その際に DOM ツリーは深さ優先で探索される.$tag_i \in \{T_b, T_l\}$ となる $tag_i$ を検出した場合, その $tag_i$ にセグメンテーションで使用するためのマークを付与する. ただし最小の単位のセグメントを検出するため,$T_b$ タイプに関しては $check\_nest$ 関数を用いて内部に $T_b$ タイプのタグがないかを確認する.$T_b$ タイプのタグが存在するならば $tag_i$ における処理を終了し, 次の DOM ツリーのタグ要素を $tag_i$ に代入する. 存在しなければ $tag_i$ にマークを付与する. この際に使用される $check\_nest$ 関数とは,$tag_i$ を `root` とする部分木の中に $T_b$ タイプのタグがないかを確認し, 存在すれば YES を, しなければ NO を返す関数である. セグメンテーションに使用するタグを示すマークは, 各タグに属性として与えている. マークが付与された $tag_i$ に対しては, さらに $tag_i$ のセグメントタイプを示す $ST(tag_i)$ 変数に, セグメントタイプ $(T_b, T_l)$ を与える. この作業を DOM ツリーに存在する全てのタグ要素に対して行い, セグメンテーションに使用するタグ要素を全て検出する.

検出ステップが終了したら, 次に挿入ステップを開始する. 挿入ステップでは検出ステップによって検出された $tag_i$ に対して, 実際にセグメントの開始と終了を示すキューを挿入する. それらのキューは,HMTL ソースに影響を与えないようにコメントノードとして DOM ツリーに挿入される. このステップにおいても DOM ツリーは深さ優先で探索され, 各タグ要素が $tag_i$ 変数に代入される. システムは各 $tag_i$ の属性をチェックし, セグメント属性を持った $tag_i$ に対してキューの挿入処理を行う. まず $state$ 変数を確認し, 終了していないセグメントがないかを確認する. 存在するならば,$tag_i$ の直前に以前のセグメントの終了キューを挿入する. 次に $ST(tag_i)$ を確認し,$ST(tag_i)$ が $T_b$ であれば $tag_i$ の直前と直後に開始キューと終了キューをそれぞれ挿入し,$state$ に close を代入する. また $ST(tag_i)$ が $T_l$ であれば,$tag_i$ の直前に開始キューを挿入し,$state$ に open を代入する. これらの作業を DOM ツリーに存在する全てのタグ要素に対して行うことによって,HTML ソース上に各セグメントの情報が付与される.

図 4.1, 図 4.2, 図 4.3 は DOM を用いてセグメンテーションする際の様子を図示したものである. この DOM ツリーに対してアルゴリズムを適用した場合, システムがどのような動きをするかを説明する.

まず検出ステップにおいて,`root` の`<html>`タグが $tag_i$ に代入される.`<html>`タグはセグメントとして使用するタグではないため, 次のタグ要素である`<body>`タグが $tag_i$ に代入され, 同様の理由で次のタグ要素である`<h1>`タグが $tag_i$ に代入される.`<h1>`タグは $tag_i \in \{T_b, T_l\}$ を満たすため,`<h1>`タグにはセグメンテーションに利用することを示す seg 属性が与えられ,$ST(T_i)$ には $T_l$ が代入される. そして次のタグである`<font>`タグが,$T_i$ に代入される. これらの処理を繰り返し,`<hr>`タグに対しても seg 属性が与えられ,$ST(T_i)$ には $T_l$ が代入される.

`<table>`タグが $tag_i$ に代入されると, まず $check\_nest$ 関数を利用し, その下の部分木に $T_b$ タイプのタグがないかをチェックする. そのようなタグが存在しないので,`<table>`タグに対して seg 属性が与えられ,$ST(T_i)$ には $T_b$ が代入される. そして次の`<table>`タグが $tag_i$

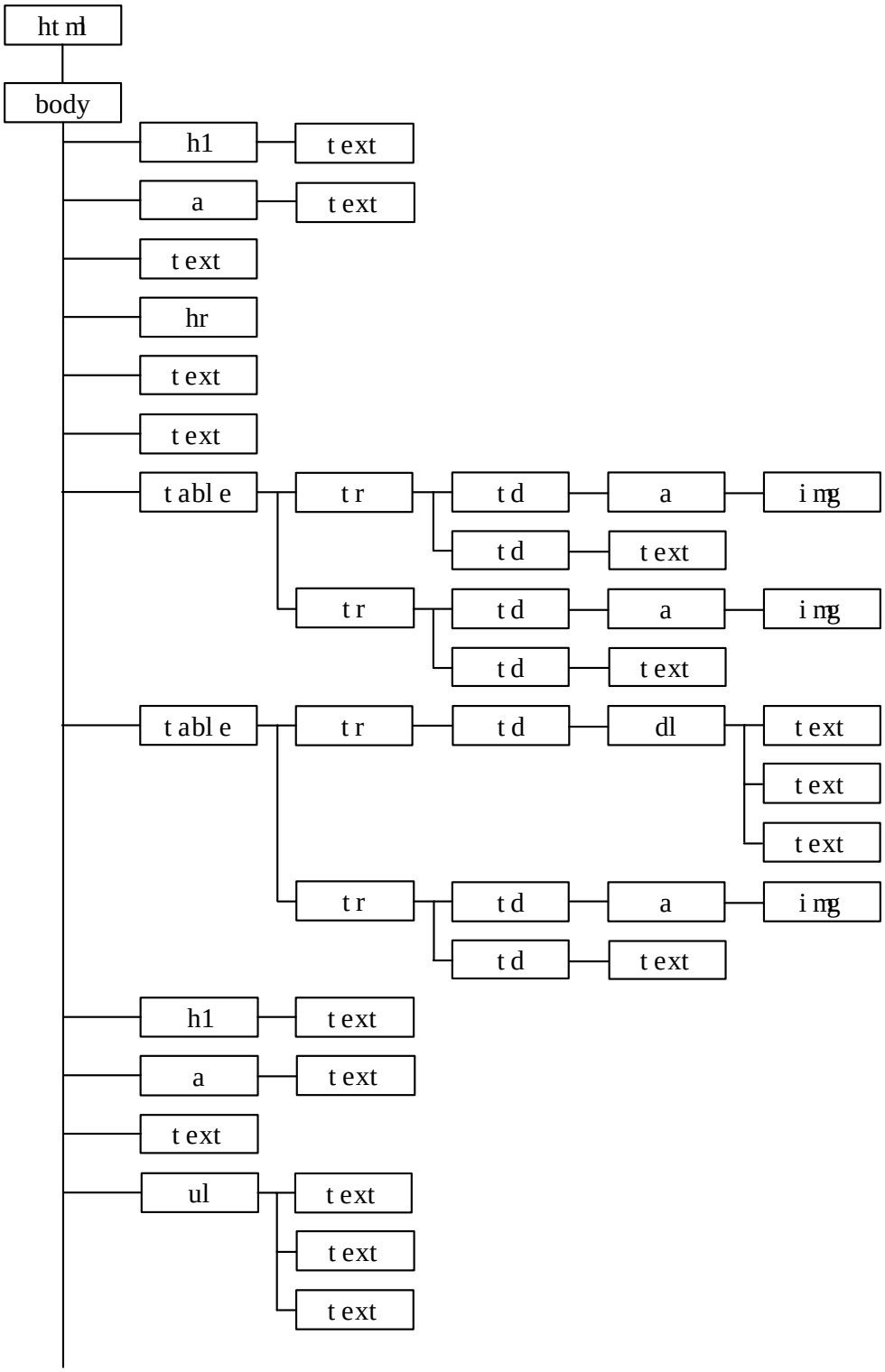


図 4.1: セグメンテーションの様子 (DOM ツリー)

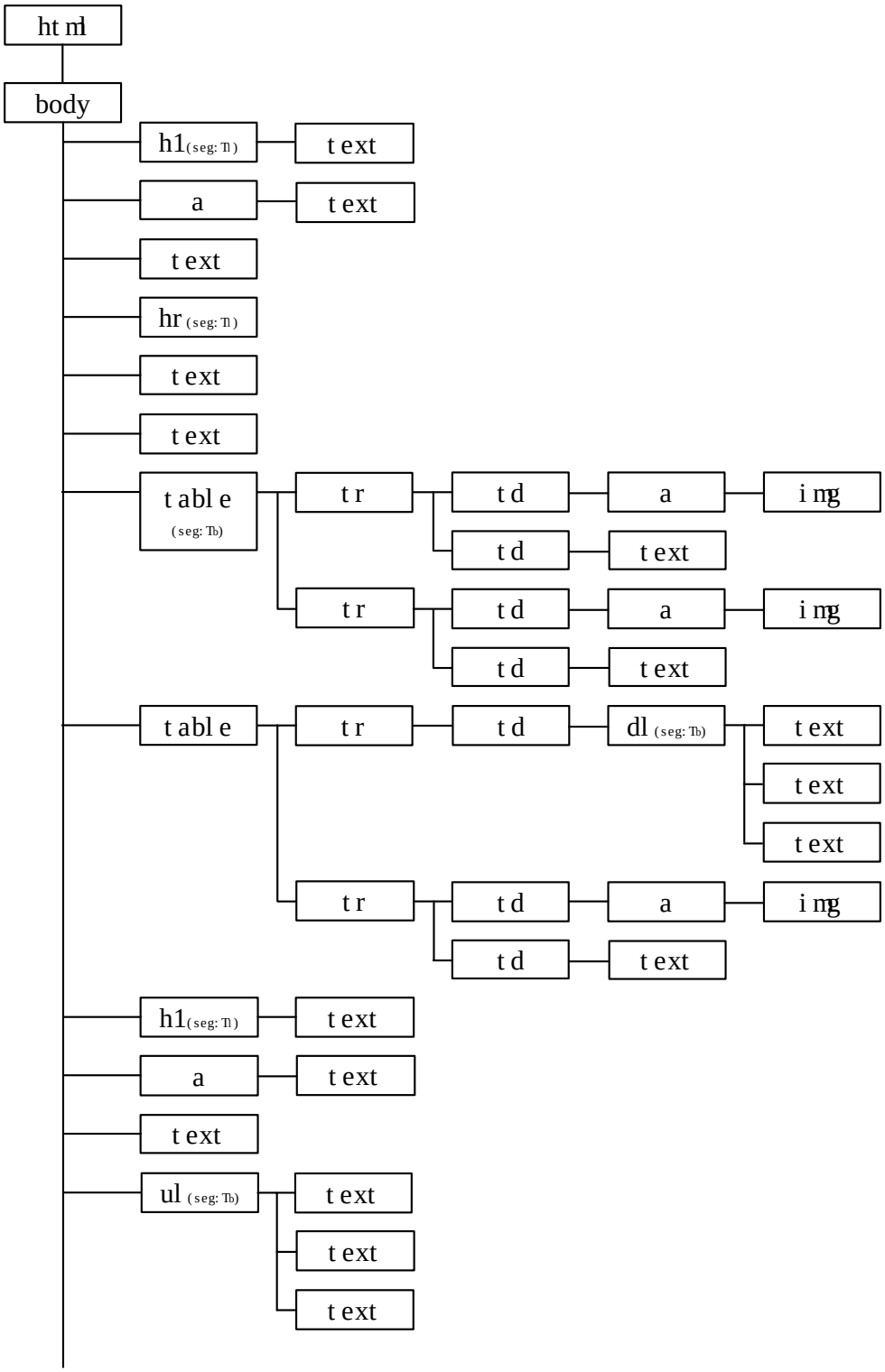


図 4.2: セグメンテーションの様子 (DOM ツリー)

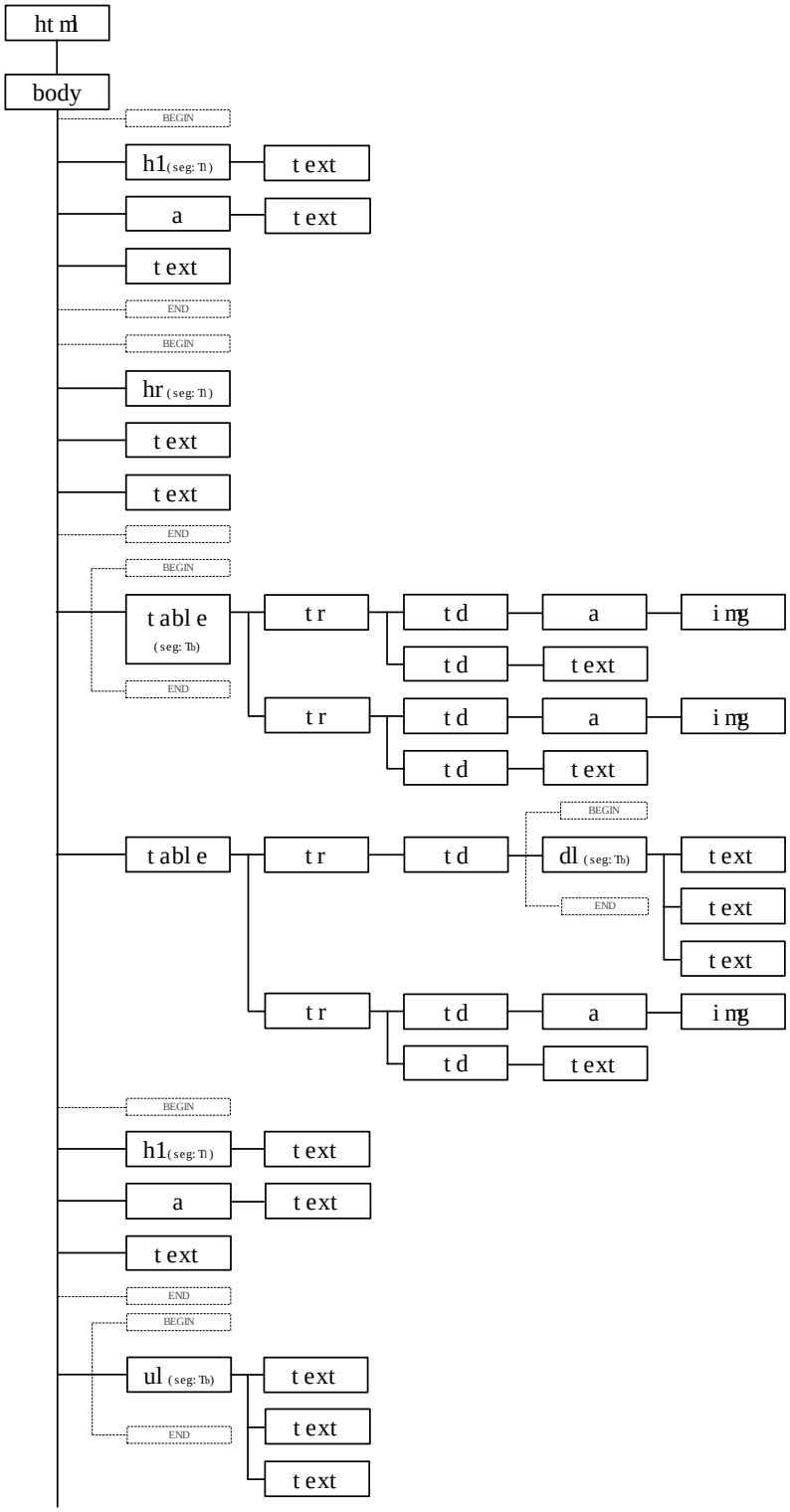


図 4.3: セグメンテーションの様子 (DOM ツリー)

に代入され, 同じように $check\_nest$ 関数を用いると $T_b$ タイプである<dl>タグが検出されるため,$tag_i$ に<dl>タグが代入されるまで木を探索する. そして $tag_i$ に<dl>タグが代入されたならば再び $check\_nest$ 関数で入れ子が存在しないか確認し,<dl>タグに seg 属性が与えられ,$ST(T_i)$ には $T_b$ が代入される.

さらに処理を進め,<h1>タグと<ul>タグに対して seg 属性を付加し, 各 $ST(T_i)$ に $T_l$,$T_b$ をそれぞれ代入する. これらの作業によって, セグメンテーションに利用できる<h1>タグ 2 個と<hr>,<table>,<dl>,<ul>タグを各 1 個ずつ検出することができる. それらのタグに seg 属性を与えたものが図 4.2 である.

次に挿入ステップの処理を行う. 検出ステップと同様に<html>タグ,<body>タグと探索し,seg 属性を持つ<h1>タグが $tag_i$ に代入されると, システムはまず $state$ を確認する.$state$ は初期値の close なので, システムは $ST(T_i)$ を確認し,<h1>タグの直前にセグメントの開始キューを挿入する. そして $state$ を open に変え, 次のタグを $tag_i$ に代入する. そして次の seg 属性を持つ<hr>タグが $tag_i$ に代入されると,$state$ の値は open なので, システムはまず先ほどのセグメントの終了を示すキューを挿入する. そして, 先ほどと同じようにタグの直前にセグメントの開始キューを挿入する. この時,$state$ の値は open のままである.

続く<table>タグの場合でも, やはりまず先ほどのセグメントの終了キューを挿入する. そして $ST(T_i)$ が $T_b$ なので,<table>タグの直前と直後に開始キューと終了キューを挿入し,$state$ を close に変える. これらの作業を続く<dl>タグや<h1>タグ,<ul>タグにおいても行う. このようにして, システムは HTML ソースにセグメントの開始と終了の情報を付与し, 図 4.3 で示されているような DOM ツリーを作成する. この DOM ツリーを深さ優先で探索していくことで, セグメントを検出することができる.

### 4.1.3 セグメントの分割

DOM 構造を用いて大まかなセグメントを検出したら, 次にセグメントの分割を行う. これは最初に検出したセグメントがしばしば大き過ぎる場合が存在するためである. 分割の手法として, 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がある.

#### イメージによるセグメント分割

DOM によるセグメント検出を行うと, 図 4.4 中で枠線で囲まれた部分がセグメントとして検出される. しかしセグメントの定義に沿って考えるならば,「オリジナル・ベストセグメント」と「レストランランキング情報」をひとつのセグメントとして,「らくらく幹事さん」と「宴会・パーティを専門スタッフがサポート」を別のひとつのセグメントとして検出すべきである. 同様にその他のテキストに関しても適切な大きさのセグメントとして検出しなければならない. 図 4.4 のように同じイメージが繰り返して出現するセグメントは, イメージを箇条書きのアイテムとして使用している場合が多い. 本研究ではこのようなセグメントに対してイメージを用いた分割手法を適用し, セグメントを適切な大きさに分

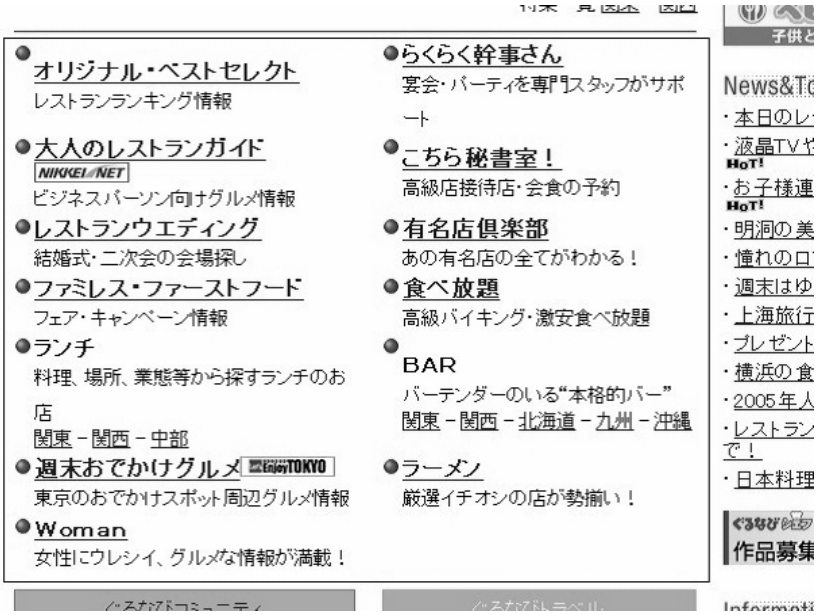


図 4.4: DOM で検出されるセグメント

(ぐるなびより抜粋)

割する.

イメージによる分割とはセグメント内に存在するイメージのうち最多のイメージを検出し, そのイメージを境界とみなすことによって, より小さなセグメントに分割する手法である. 手法の手順は以下のとおりである.

1. 各初期セグメント内に存在するイメージの各個数を数える
   - 最多のイメージが 5 個以上であれば, 分割システムを開始する
   - 最多のイメージが 4 個未満であれば, 分割しない
2. 最多のイメージをセグメントの境界とみなし, イメージの最初のものから次のイメージまでをひとつのセグメントとする
3. 分割セグメントが 2 個以上のテキストを含んでいるかを検査する
   - 満たしていれば, イメージの前にセグメントの終了と開始を示すキューを配置する
   - 満たしていなければ, 次のイメージを先頭とするセグメント分割処理を行う
4. 2,3 の作業を繰り返す

5. 3 の作業において分割セグメントが作成された場合, 初期セグメントのキューを削除する

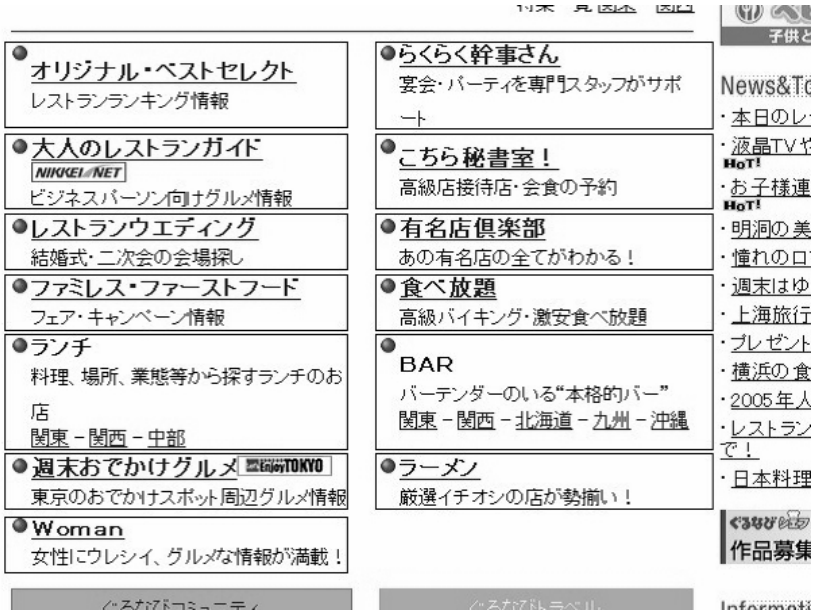


図 4.5: 正解セグメント

最初に述べた図 4.4 のセグメントの場合, イメージによる分割手法を適用すると, まず見出しの前にある丸のイメージが最多のイメージとして検出される. そして最初のイメージから次のイメージまでの間にあるテキスト (「オリジナル・ベストセレクト」と「レストランランキング情報」) をひとつのセグメントとする. そしてこのセグメントが定義に合っているかを検査し, 合っているのでセグメントとして HTML ソースにセグメントの開始と終了を示すキューを配置する. この手順を繰り返すことによって, それ以後の 12 個のセグメントも検出することができ,DOM を用いたセグメンテーションではひとつの大きなセグメントとして検出されていたものが, 図 4.5 に示されているような 13 個の適切な大きさのセグメントに分割される.

### テーブル内の部分木を用いたセグメント分割

図 4.6 で示されているセグメントは DOM に基づいて検出されたものである. このセグメントも「便利なツール」と「ビジネス」という見出しを持った二つのセグメントに分割する必要がある. しかし, これらのセグメントにはイメージが使用されていないため, 前述のイメージによるセグメント分割手法は適用できない.

イメージを含まないセグメントを分割する手法として, 部分木の類似度を利用したセグメント分割手法を導入した. これは`<table>`タグ内の`<tr>`タグ, もしくは`<td>`タグを `root`

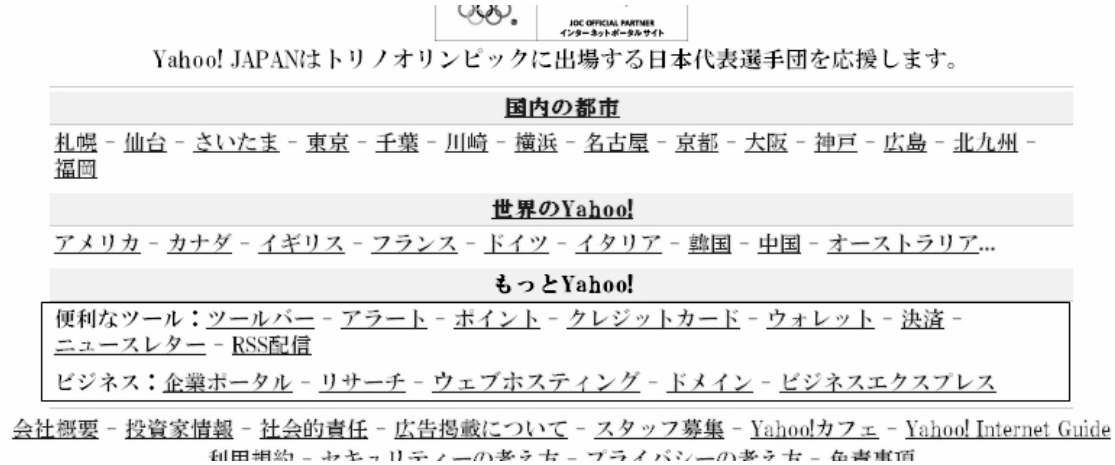


図 4.6: DOM で検出されるセグメント

(Yahoo!より抜粋)

とする DOM の部分木を取り出し, それらの類似度が閾値以上であれば `root` の単位で分割する手法である. システムの手順は以下のとおりである.

1. `<table>`タグによってセグメンテーションされたセグメントを検出する
2. `<td>`タグを `root` とするタグの部分木を作成
3. 作成した部分木から共通部分木を取り出す
4. 共通部分木と各テキストの部分木の最小のものと比較
5. 4 の類似度が閾値以上であれば各部分木をひとつのセグメントとして分割する
6. 4 の類似度が閾値未満であれば`<tr>`タグを `root` とするタグの部分木を作成し, 同じ処理を繰り返す

共通部分木の抽出にはデータマイニングツールの freqt[9] を用い, 全ての部分木に共通する木を抽出する. それらの共通部分木のうち, ノード数が最大の部分木を用いて類似度を計算する. ここで述べている類似度とは,

$$\frac{\text{最大の共通部分木のノード数}}{\text{取り出した部分木のうち最小の木のノード数}}$$

として定義されたものである. 閾値に関しては 0.5 に設定した.

例として図 4.6 と図 4.8 のページ例を示す.DOM を用いたセグメンテーションでは, テーブル部分に囲まれたセグメントとして図 4.6 のようなセグメントが検出される. しかしセグメントの定義で考えると, 図 4.8 に示されているようなセグメントが適切なセグメント

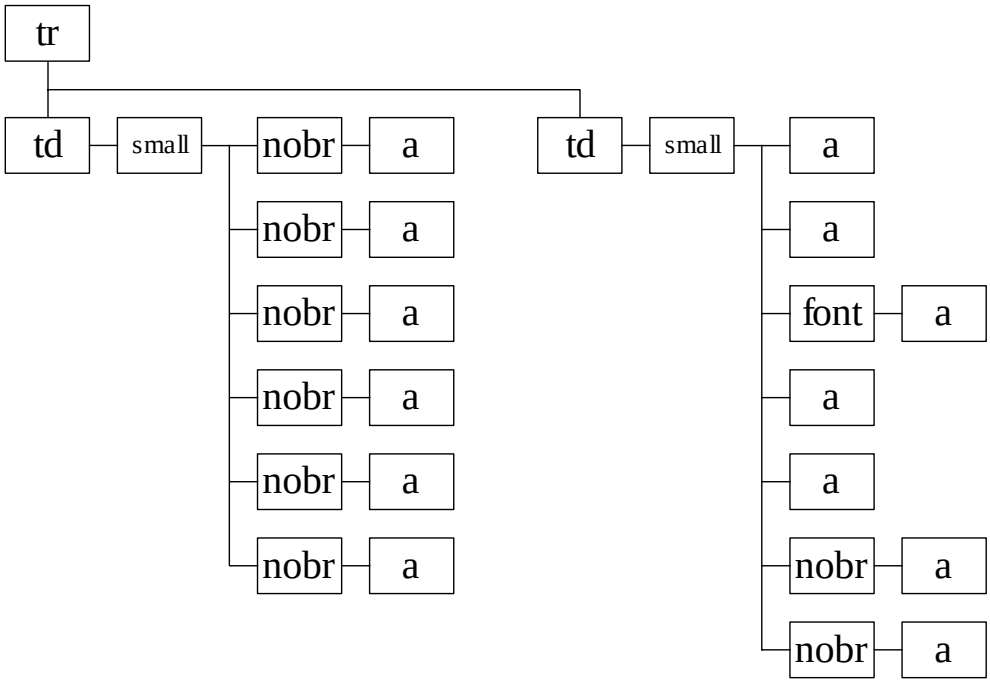


図 4.7: DOM ツリー

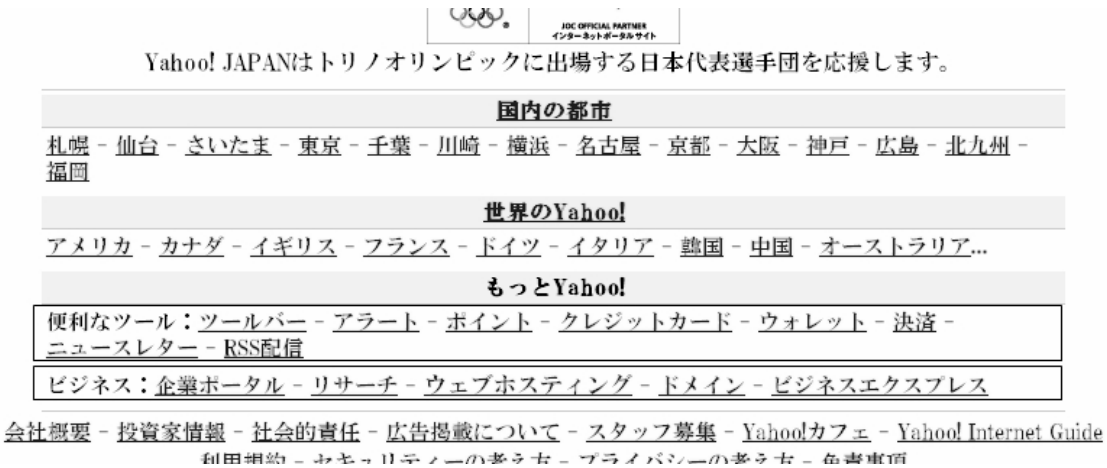


図 4.8: 分割されたセグメント

であるといえる. このような場合において, テーブル内の部分木を利用した分割が行われる. このテーブル内の DOM 構造は図 4.7 のようになっている.`<td>`タグを `root` とした部分木を作成すると,

```
(td(small(nobr(a))(nobr(a))(nobr(a))(nobr(a))(nobr(a))(nobr(a))))
(td(small(a)(a)(font(a))(a)(a)(nobr(a))(nobr(a))))
```

の形で出力される.freqt を用い, この二つの部分木の最大共通部分木を取り出すと,

```
(td(small(nobr(a))(nobr(a))))
```

が取り出される. 従ってこの二つの部分木の類似度は

$$\frac{\text{共通部分木のノード数}}{\text{取り出した部分木のうち最小の木のノード数}} = \frac{6}{12} = 0.5$$

と算出され, 類似度が閾値の 0.5 以上であるので, 各部分木はひとつのセグメントとして分割される. テーブル全体で定義されたひとつの大きなセグメントが, 図 4.8 に示されているような適切な大きさのセグメントに分割される.

この分割手法は, テーブル内の共通部分木を取り出して類似度を判定することによって, テーブル内に存在する繰り返し構造を検出する. 従って本手法は南野らの研究 [3] で示されていた繰り返し構造を利用したものである.

### 4.1.4 セグメントのマージ

ここまでに検出されたセグメントは最小の単位のセグメントであり, リストタグのようなヘッダ部分を含まないセグメントが存在する. しかし本研究の提案システムでは, ヘッダ部分はセグメントの内容を示す重要な手がかりとみなし, ヘッダ部分が存在するセグメントではヘッダ部分の読み上げを行う. 従って, セグメントとそのヘッダが 1 つのセグメントになっていないとき, それらをマージする必要が生じる. 本研究で用いた手法は, 次のようなものである.

1. ヘッダ候補となるテキストを検出する
2. そのテキストの後方に隣接するセグメントがあるかどうかを探索する
   - ヘッダ候補とセグメントの間で次の条件を満たす場合, それらは隣接しているとみなす
     - テキストが存在しない
     - `alt` 属性を持つイメージが存在しない
     - `alt` 属性を持たないイメージが 3 個以下
3. セグメントが存在すれば, テキストとセグメントをマージする

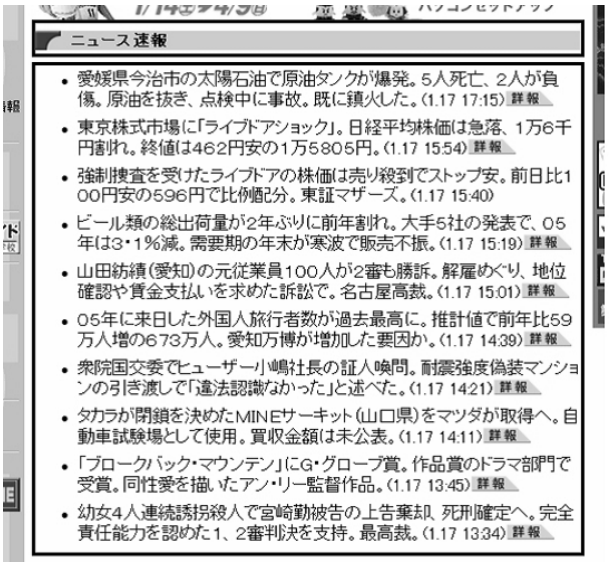


図 4.9: マージ例 (処理前)

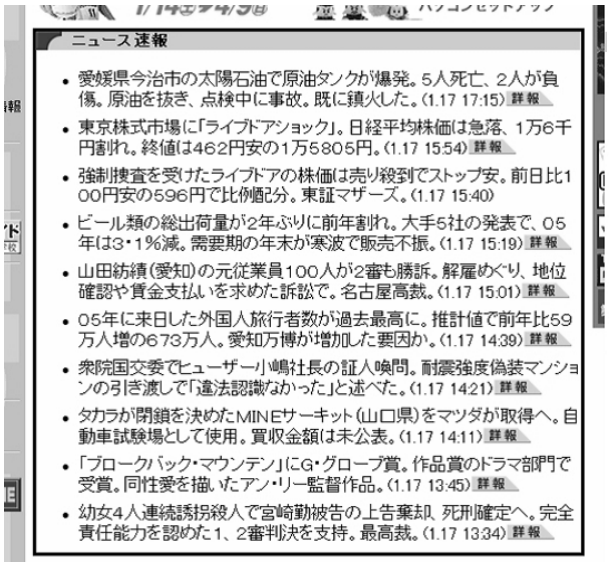


図 4.10: マージ例 (処理後)

(中日新聞ホームページより抜粋)

ヘッダ候補となるテキストを検出するために以下の特徴を利用している.

- h1, h2, h3, h4, h5, h6
- b
- thead

マージの例として, 図 4.9 と 4.10 のウェブページを示す. 図 4.9 で示されているセグメントは DOM-Based セグメンテーションの際に`<ul>`タグにより検出されたセグメントである. この例を見ても分かるとおり, リストタグによって箇条書きにされているものはヘッダを含まない. 従って, ヘッダとなる「ニュース速報」のテキストと隣接するセグメントをマージし, 図 4.10 に示されているようなセグメントにしなければならない. この例の場合, 「ニュース速報」というテキストは`<b>`タグによって修飾されているため, システムはこのテキストをヘッダ候補として認識する. 次に隣接するセグメントの探索を行い,`<ul>`タグによるセグメントを検出する. その際, ヘッダ候補とセグメントの間にテキストやイメージが存在しないかチェックし, 存在しないのでシステムはヘッダ候補とセグメントをマージし, 図 4.10 のようなセグメントを新たに作成する.

## 4.2 実験

### 4.2.1 評価方法

評価は人手で作成した正解セグメントとシステムによって検出されたセグメントとの比較によって行った. セグメント情報が付与された HTML ソースよりテキストとイメージのみを取り出し, システムによって付与されたセグメントの位置が正解セグメントの位置と一致しているかを判定した. これは音声ブラウザが読み上げるのはテキストとイメージであるためで, このようにすることによって, セグメントのキューが挿入された位置がタグの間で前後したとしても, 正しく判定することができる. さらにイメージに関しては `alt` 属性を持つイメージのみを取り出している. これは音声による読み上げの際に `alt` 属性を持たないイメージはユーザに対して有用な情報を与えることはできないという仮定に基づいている. 判定はセグメントの開始位置と終了位置が一致しているかどうかを判定した完全一致評価と, 開始位置のみの一致を判定する半一致評価を行っている. 完全一致だけでなく半一致評価を行った理由は, 本システムの主な目的がページの構造化ではなく音声ブラウザにおけるナビゲーションであるためである. つまりページの読み上げをナビゲートする際にはセグメントの開始位置が重要であり, 終了位置は他のセグメントと重複していない限りは流動的であっても良いという考えに基づいている.

### 4.2.2 使用データ

実験はクローズテストとオープンテストの 2 種類を実施し, 各テストは 20 のウェブページを用いて行った.

クローズテストに用いたウェブページの内訳は, ポータルサイト 4 ページ, ニュースサイト 2 ページ, ショッピングサイト 3 ページ, 企業・公的機関のページ 5 ページ, 個人サイト 6 ページである. 各ページに対し正解データとしてセグメントの開始と終了箇所を示すコメントを人手で HTML ソースに挿入した. その際に作成された正解セグメントは, レンダリングされたページに基づいて判断されている. システムの作成に当たり, これらの 20 個のページに対し, 実験を行いエラー解析を行うことでシステムの改良を重ねた. 従ってこれらのウェブページはクローズドなデータといえる.

クローズテストとは別にオープンテストも実施した. オープンテストに使用したページの内訳は, ポータルサイト 9 ページ, ニュースサイト 1 ページ, 企業・公的機関のページ 4 ページ, 個人サイト 6 ページである. これらはシステム開発者ではない第三者によって収集され, 本研究のセグメントの定義に沿って正解セグメントが付されたものである. 第三者テストデータを収集することにより, システムに有利なページだけを偏って集められることがないようにした.

### 4.2.3 実験結果

#### クローズテスト

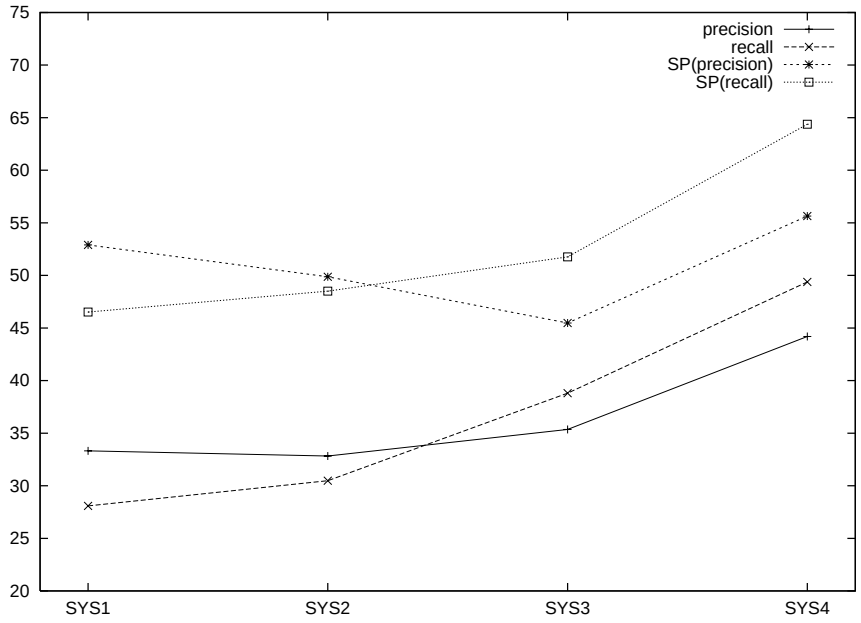


図 4.11: セグメンテーション結果 (クローズテスト)

表 4.1: セグメンテーション結果 (クローズテスト)

| | **seg** | **precision** | **recall** | **SP(precision)** | **SP(recall)** | **cross** |
|---|---|---|---|---|---|---|
| SYS1 | 7.9 | 33.32% | 28.09% | 52.90% | 46.52% | 0.55% |
| SYS2 | 8.5 | 32.83% | 30.48% | 49.87% | 48.51% | 1.20% |
| SYS3 | 10.4 | 35.36% | 38.80% | 45.47% | 51.76% | 0.80% |
| SYS4 | 11.8 | 44.19% | 49.38% | 55.64% | 64.38% | 1.43% |

**オープンテスト**

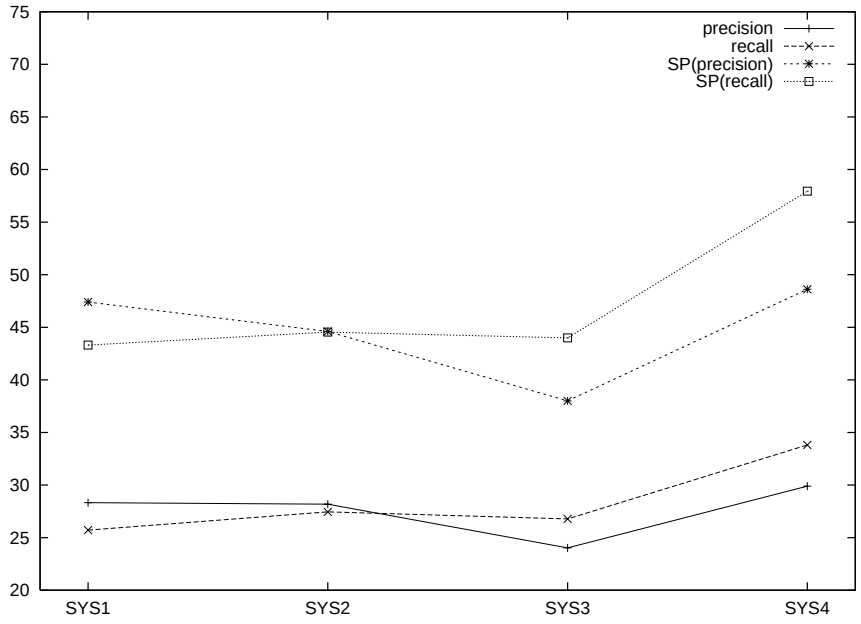


図 4.12: セグメンテーション結果 (オープンテスト)

表 4.2: セグメンテーション結果 (オープンテスト)

| | **seg** | **precision** | **recall** | **SP(precision)** | **SP(recall)** | **cross** |
|---|---|---|---|---|---|---|
| SYS1 | 5.9 | 28.33% | 25.71% | 47.40% | 43.31% | 1.22% |
| SYS2 | 6.4 | 28.18% | 27.45% | 44.60% | 44.54% | 2.08% |
| SYS3 | 6.5 | 24.03% | 26.78% | 38.00% | 43.99% | 1.73% |
| SYS4 | 7.7 | 29.89% | 33.81% | 48.61% | 57.95% | 3.35% |

図 4.11・表 4.1 はクローズテストの, 図 4.12・表 4.2 はオープンテストの各システムの precision と recall 等の 20 ページの平均を示したものである. 表の中の seg は正解したセグメント数の平均,precision は自動検出されたセグメントのうち正解セグメントと一致したものの確率であり,recall は正解セグメントのうちシステムがどれだけのセグメントを完全に検出できたかを示す確率となる. また SP とはセグメントの開始位置のみの precision と recall を算出したもの,cross とは自動検出されたセグメントのうち正解セグメントをまたいでいるものの確率を示したものであり, 自動検出セグメントのエラー率となる.

左の行に書かれている SYS は各手法を導入したセグメンテーションシステムの種類を指している.SYS1 は最も単純な DOM-Based のセグメンテーションシステムである (4.1.2 項参照).SYS2 は SYS1 にイメージ分割手法 (4.1.3 項の「イメージによるセグメント分割」参照) を,SYS3 はイメージ分割に加えて, テーブル分割手法 (4.1.3 項「テーブル内の部分木を用いたセグメント分割」参照) を付加したものである. 最後の SYS4 は SYS3 にマージシステム (4.1.4 項参照) を付加したもので, 本研究で作成したシステムの最終版となる.

実験結果の詳細は付録を参照していただきたい. 付録では 20 の個々のウェブページにおける各システムの結果と, ページのタイプ別に結果をまとめた表を載せた.

### 4.2.4 考察

実験結果より本研究で作成した自動セグメンテーションシステムが一定の成果を挙げていることが確認できた. 表 4.1 を見ると, クローズテストでは平均で約半数の 49%のセグメントを完全に取り出すことができている. この 49%という数字はあまり良いものとは言えないが, 開始位置だけで言えば 64%のセグメントが一致しているため, これらのセグメントに関しては, 終了位置をより適切に検出することのできるセグメンテーション手法を適用することによって完全に検出できるようになる. 一方,precision の結果を見るとクローズテストでは半一致の結果が, オープンテストでは完全一致と半一致のどちらの結果も分割手法を導入することで SYS1 の結果より低くなっている. これは分割によって, 検出されるセグメントの総数が増加したことが原因と考えられる. しかし,cross は 2%以下に収まっていることからシステムが検出したセグメントが複数の正解セグメントにまたがっていることはほとんどなく, 更なる手法の改良によって本システムの精度を向上させることができると考えられる.

SYS1 から SYS4 までの結果の推移を見ると, クローズテストにおいては単純な DOM-Based の SYS1 に, イメージ分割, テーブル分割, マージの各手法を導入することによって精度が向上していることが確認できる. 従って本研究で作成したセグメンテーション手法の有効性は明らかである. しかしオープンテストでは,SYS3 の結果が SYS2 の結果より低下していることが確認された. ここからテーブル分割手法の改善が必要なことが分かる. 本手法では行っていないが, 各テキストにかかるタグの出現順序に関する情報に関して類似度を用いる際には, 南野らの研究で行われているようなタグをテキストの属性として置き換えることを行う必要もあり, 将来的にはそのような手法も加えていかなければならない

と思われる. また分割手法の導入によって precision が低下することを防ぐため, 分割されたセグメントが適切かを判定する際に分割セグメントに特化した判定基準を設ける必要がある.

また, 本研究で用いた正解セグメントは入れ子を許しているが, システムが検出するセグメントは入れ子を許していない. 従って今後は現在までに開発したシステムの再帰的適用を考慮に入れ, より適切なセグメントの検出を目指す.

その他にも, 本研究において作成したシステムは HTML ソースにあるタグ情報のみを使用してセグメンテーションしているが, 今後はテキストの言語情報やコンテンツ間のレイアウト情報といった現在使用していない情報を利用し, ヘッダ部分の特定や現行の DOM-Based のシステムでは検出することのできないセグメントの検出を行っていきたい. またウェブページのセグメンテーションは本研究の目的である視覚障害者支援だけではなく, 情報抽出や情報検索といった様々なシステムに有用なものである. 従ってさらなる精度の向上や汎用性の向上を図っていきたいと考えている.

# 第5章　リンクの参照位置の同定

本章では, ユーザの目的位置から読み上げを開始するシステムの中で用いるリンクの参照位置を同定する手法を紹介する. リンクの参照位置とは, ユーザがリンクをクリックした際に目的としているリンク先ページ内にあるテキストの位置を指す.

1 章において, 音声ブラウザを用いてウェブブラウジングを行う際にユーザの手間が増大する原因として, 音声ブラウザは HTML ソースの中の出現順序に従って単純に読み上げを行っていることを挙げた. この問題点はユーザがページの内容を閲覧するためにリンクをクリックし, 新たなページを開く際にも手間を発生させる原因となる. つまり図 5.1 のようなニュースページにおいてユーザが記事をクリックした場合, 図 5.2 のページが開かれるが, そのページの読み上げは再び一番上から開始される. 画面から情報を得ることのできる人であれば, レンダリングされたページを見て即座に記事の部分を見つけ, 記事を読むことができる. しかし音声によって情報を得るユーザは

「違法性の認識ない」証人喚問でヒューザー小嶋社長

という記事のタイトル部分とそれに続く内容部分までたどり着くまでに, ページの上部にある広告やコンテンツを全て読み上げなければならない.

上記の問題点を解消するために, 本研究ではユーザの目的位置から読み上げを開始するシステムの開発に取り組んできた. 本研究で考案したシステムの手順は次のようなものである.

1. ユーザがリンクをクリックする
2. クリックしたアンカー文字列から, ページ内位置の同定を行う
3. ユーザに対して, リンク先ページに目的位置と思われる箇所を検出したことを通知する
4. ユーザが望めば, 検出された文字列を持つテキストと隣接するテキストを読み上げる

このシステムを実現するために, 本研究ではリンクの参照位置の同定を行う手法を開発した. 本研究における詳しい手法の説明は 5.1 節で述べる.


図 5.1: ニュースページ例 (asahi.com より抜粋)


図 5.2: リンク先の記事ページ例 (asahi.com より抜粋)

## 5.1 提案手法

リンクが指すページ内位置の同定を行う手法として, 本研究で作成したシステムはリンク元ページのアンカー文字列とリンク先ページのテキストとの間でパターンマッチを行っている. システムの手順は以下のとおりである.

1. アンカー文字列とリンク先のウェブページを読み込む
2. リンク先ページ内のテキストを取り出す
3. アンカー文字列とリンク先のページ内のテキストの長さを調べる
   - ページ内テキスト長が 5 文字未満であればマッチを行わない
     (短すぎるテキストがマッチすることを防ぐため)
4. テキスト長が短いものをパターンとして, 長い側のテキストとマッチさせる
   - マッチした場合, テキスト番号を格納する
5. リンク先ページ内にある各テキストに対して 3,4 の手順を繰り返す
6. 格納したテキスト番号に従って, テキストを読み上げる

システムは, リンク先ページの HTML ソース中にある全てのテキストと `alt` 属性を持つイメージに対して, パターンマッチを行う. 従って本システムは各ページに対して通常複数のリンクの参照位置を検出することになる. 完全一致ではなくパターンマッチを使用することによって検出されるテキストは増加するが, このシステムは一回で回答を出すことよりも確実に参照位置を同定することが要求されるため, テキスト照合に成功する可能性の高いパターンマッチを使用した.

パターンマッチを使用することによって, 完全一致では検出することのできないリンクの参照位置を同定できた例を図 5.3, 図 5.4 に示した. この例では, リンク元のページから得られるアンカー文字列は

「■第 74 回日本音楽コンクール」

である. しかしリンク先ページの参照位置には「第 74 回日本音楽コンクール」となっている. 従ってアンカー文字列の中に「■」が存在するため, 完全一致では同定することはできない.「ボイス・サーフィン」の記事本文アクセス機能もアンカー文字列の前方部分からマッチする文字列を探していると思われ,「■」のないリンクの参照位置から記事本文を検出することができなかった. しかし本手法では, これらのページにおいても正しくリンクの参照位置を同定することができる.

もうひとつの例として図 5.5, 図 5.6 を示した. ここで示した図 5.5 のページでクリックしたアンカー文字列は

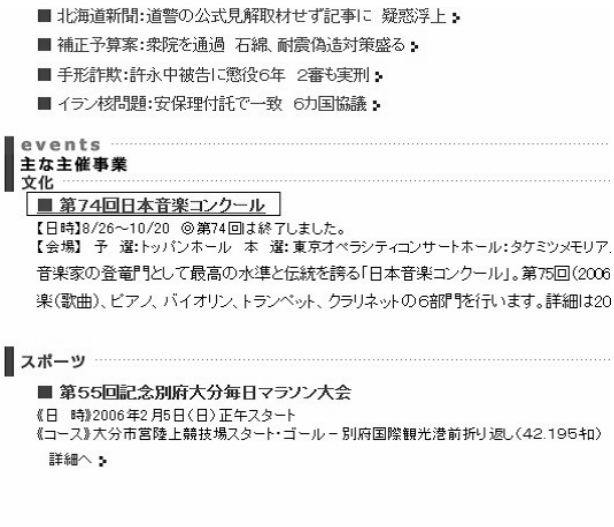


図 5.3: 後方一致例 (リンク元ページ)

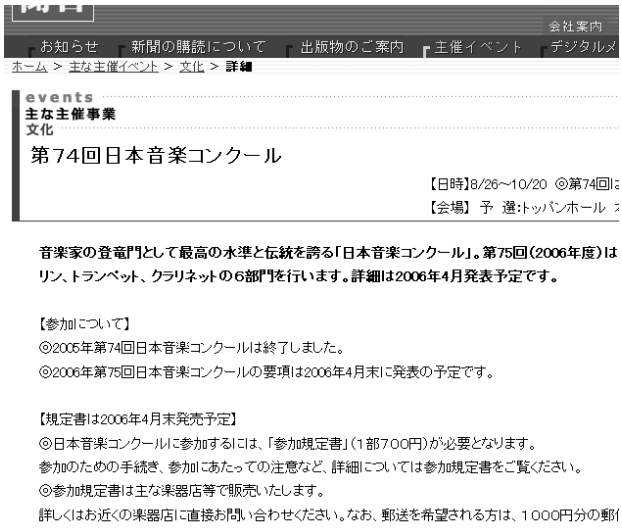


図 5.4: 後方一致例 (リンク先ページ)

(毎日新聞ホームページより抜粋)

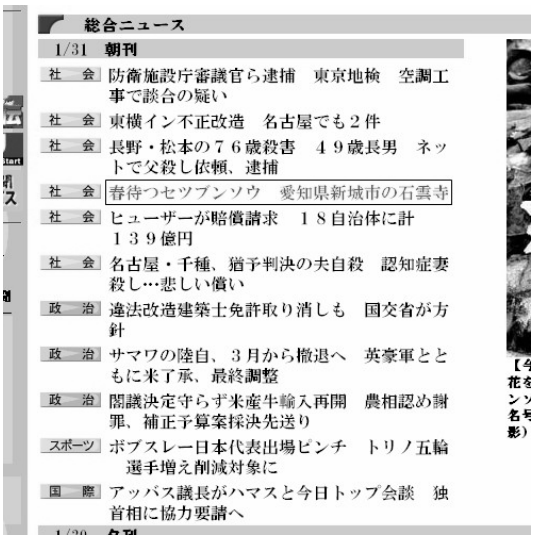


図 5.5: 二行分割例 (リンク元ページ)


図 5.6: 二行分割例 (リンク先ページ)

(中日新聞ホームページより抜粋)

「春待つセツブンソウ 愛知県新城市の石雲寺」

である. そのリンク先となっているページを示した図 5.6 を見ると, 参照位置となるべきテキストが二行に分かれていることが分かる. この場合も完全一致では検出することはできない. しかし本システムではアンカー文字列よりリンク先ページ内のテキストが大きい場合には, リンク先テキストをパターンとして, アンカー文字列に対してマッチを行っている. その結果, 図 5.5, 図 5.6 のようなページにおいてもリンク参照位置を正しく同定することができる.

## 5.2 実験

### 5.2.1 評価方法

本手法の評価方法として, アンカー文字列を用いてリンク先ページから目的となる文字列を取り出せるかどうかを判定した. 実験方法としては, 各リンク先ページの目的位置となるテキストの番号を正解データとして作成しておき, システムが検出したテキストの番号と比較し, 正しく取り出せたかどうかを判定した.

### 5.2.2 使用データ

実験用データとして 20 ページから各 5 個のリンクを選択し, 計 100 ページにおいてシステムの精度を実験した. ページの種類はニュースサイト 9 個, ポータルサイト 3 個, 企業, 公的機関のページ 5 個, ショッピングサイト 2 個, 個人サイト 1 個である. ニュースサイトを多く収集したのは, このシステムは主にニュースサイトのようなページを対象として開発したためである.

### 5.2.3 実験結果

表 5.1 は実験結果を示した表である. 実験データ 100 ページの平均を表した.detect は各ページにおいて参照位置として検出されたテキストの数の平均,precision と recall は次の様な式で算出した.

$$precision = \frac{\text{正解した参照位置の全テキスト数}}{\text{検出された全テキスト数}}$$

$$recall = \frac{\text{正解した参照位置テキスト数}}{\text{正解参照位置テキスト数}}$$

この式を見ても分かるように,recall は 0 か 1 のどちらかになる.

表 5.2 は, ウェブページの目的位置までたどり着くのに読み上げが必要なテキスト要素数の平均を示したものである.NON は従来のリンク先のページを HTML ソースの出現順に単

表 5.1: 実験結果 (同定)

| detect | precision | recall |
|---|---|---|
| 3.70 | 18.11% | 67.00% |

表 5.2: 実験結果 (テキスト要素数)

| ALL Pages | | | Identifiable Pages | | |
|---|---|---|---|---|---|
| page num | NON | JUMP | page num | NON | JUMP |
| 100 | 68.23 | 18.2 | 67 | 76.13 | 1.46 |

表 5.3: 実験結果 (ページタイプ別)

| page type | page num | precision | recall |
|---|---|---|---|
| news | 9 | 43.18% | 84.44% |
| portal | 3 | 17.95% | 46.67% |
| business | 5 | 38.89% | 56.00% |
| shopping | 2 | 3.92% | 80.00% |
| personal | 1 | 0.00% | 0.00% |

純にたどっていったときのテキスト数を,JUMP はシステムが検出したテキストのうち, 何個目の検出テキストが目的位置テキストであったかをカウントしたものである.ALL Pages は全ページにおける結果であるが, 本システムで参照位置が同定できなかった 33 ページに関しては NON と同じステップ数がかかることとして算出した. もう一方の Identifiable Pages は参照位置を同定できた 67 ページの結果を示したものである.

表 5.3 は, ウェブページのタイプごとの実験結果の平均を示したものである.news はニュースサイト,portal はポータルサイト,business は企業公的機関サイト,shopping はショッピングサイト,personal は個人サイトを表している.

### 5.2.4 考察

表 5.1 より, 単純なパターンマッチの手法を用いることで 67%の性能を得ることができた.precision に関しては, システムを recall 重視で作成したため 18%と低い数字になっている. しかし検出された 5 個のうち 1 個が正解しているならば視覚障害者支援の用途に十分耐えうるものであると思われる.

表 5.2 より, 本システムを用いることでリンクの正しい参照位置までに読み上げるテキスト要素数を減らせることが確認できた. 本システムによって, 全く使用しない時から約 73%のテキストの読み上げの省略を行うことが出来ている. このことから本システムが

67%という精度であっても, 視覚障害者支援に有効であることが確認できた. さらに正しく位置を同定できた 67 ページについて同様の実験を行った結果, システム未使用時の参照位置までのテキスト要素数が 76.1 だったのに対し, システムを使用した場合のテキスト要素数は 1.5 となっており, 実に読み上げの 98%の省略を行っている.

ページのタイプ別に結果を示した表 5.3 を見ると, ニュースサイトタイプのページにおいて良好な結果を得ることができた. 従って本システムの当初の開発意図であるニュースページにおけるリンクの参照位置の同定は, ほぼ成功しているといえる. その他のものではショッピングサイトや企業公的機関のサイトでは比較的良い結果が得られたが, これはそれらのページにおいては, アンカー文字列と参照位置の文字列の間に言い換えが少なかったことが良い結果につながったと考えられる. 反対にポータルサイトや個人サイトにおいては良い結果が得られなかった. 特に個人サイトに関しては,1 種類のページにおいてしか実験を行っていないため, 今後更なる実験を進める必要があるが, 今回の実験ではアンカー文字列とリンク先の目的文字列の間にある言い換えがポータルサイト, 個人サイトに共通して多かったことがエラーの原因と考えられる.

このエラー原因となっている言い換えについて説明する. ここで言う言い換えとはアンカー文字列とリンク先ページ内の参照位置テキストが別の表現で表されていることを指す. 本研究ではリンクが指すページ内のテキストは同じ形で存在するという仮定の下でシステムを作成したが, アンカー文字列と参照先のテキストの間に表現の差異があると正しく同定できない場合が存在した. 実験によって判明した本システムが検出できない表現の差異は以下の 3 種類である.

- テキスト内部に文字列が付加されて生じた差異
- アンカー文字列, 参照位置文字列それぞれに文字が付加されている差異
- 同じ意味の単語に言い換えられている差異

本研究で作成したシステムではパターンマッチを導入することにより, テキストの前後に付加された文字列があったとしてもシステムはテキストを検出することができる. しかしテキスト内部に文字列が付加されるとシステムはテキストを検出することができなくなってしまう. 例として, 図 5.7 と図 5.8 のウェブページを挙げる. ここでリンク元ページから得られるアンカー文字列は,

「個人情報について」

である. しかしリンク先のページにある参照位置文字列は

「個人情報保護に関する考え方について」

となっているため, システムはこの文字列を検出することができなかった. 従って今後はテキストが内部に付加されたものも検出する手法の開発が必要となる.

またアンカー文字列とリンク先文字列それぞれに細かな言い換えがある例として, 図 5.9 と図 5.10 がある. 図 5.11 の中で

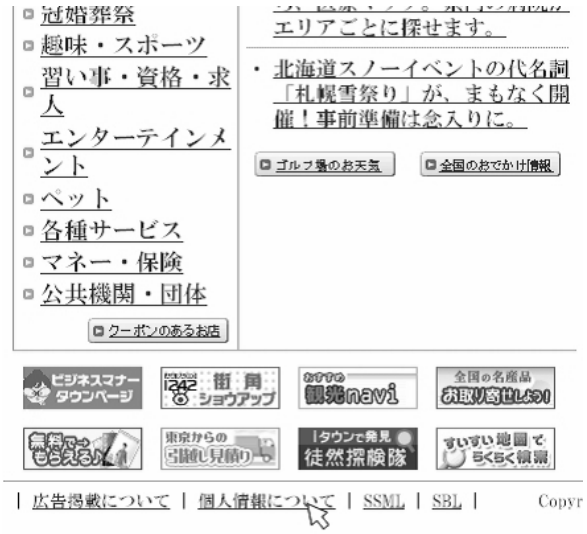


図 5.7: 言い換え例 1(リンク元ページ)

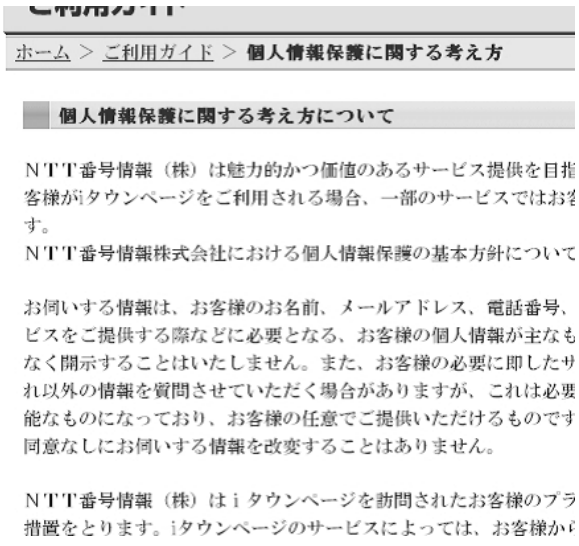


図 5.8: 言い換え例 1(リンク先ページ)

(i タウンページより抜粋)

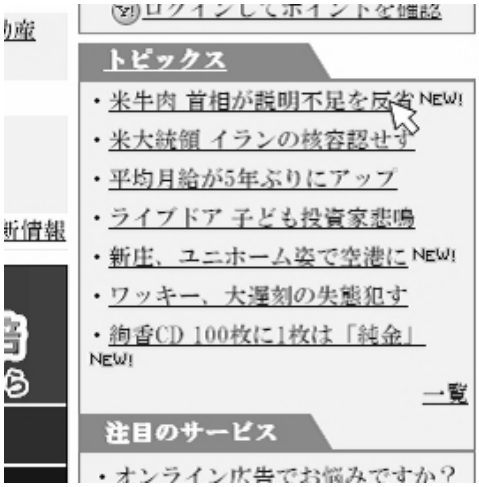


図 5.9: 言い換え例 2(リンク元ページ)

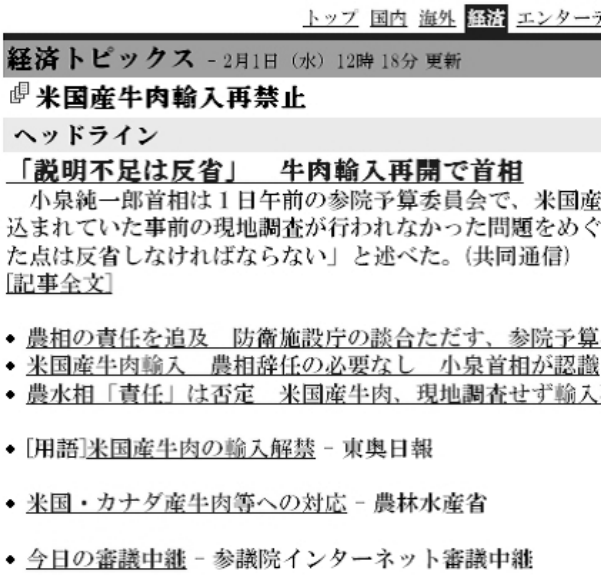


図 5.10: 言い換え例 2(リンク先ページ)

(Yahoo!より抜粋)

「米牛肉 首相が説明不足を反省」

という記事をクリックした場合, 図 5.12 のページが開かれる. しかしこのリンク先のページでは, 記事のタイトルが

「「説明不足は反省」 牛肉輸入再開で首相」

となっており, これらはパターンマッチを用いた手法では検出することができない. 従ってこれらのテキストを検出するためには, 単語ベクトルの比較を用いた手法を導入しなければならない. 単語ベクトルを用いることによって各単語の類似度が求められるため, アンカー文字列とリンク先文字列のどちらにも文字列が付加されていたとしてもシステムは正しくリンク先の参照位置を同定できると期待できる.

さらに言い換えの別の例として, 図 5.11 と図 5.12 を挙げる. 図 5.11 の中で

「JAWS V4.5(Q & A)」

という記事をクリックした場合, 図 5.12 のページが開かれる. しかしこのリンク先のページでは, 記事のタイトルが

「JAWS V4.5 のよくある質問」

となっており,「Q & A」が「よくある質問」に言い換えられている. このような問題に対処するためには, 頻出する言い換え表現を記した辞書を準備し, そこで定義されている様々な表現を用いて単語ベクトルによる比較を行う必要があると考えられる.

また実験ではシステムがリンク先ページ内で 1 個以上のテキストを検出したが, その中に正しいテキストが含まれていない事例も見られた. 誤りの原因を調査したところ, 参照位置のテキストは言い換えが行われており検出できないが, ページ内のメニューの中に記事タイトルが記載されており, システムがそれらを誤検出していたことが分かった. 図 5.13, 図 5.14 はその一例である. ここでは図 5.13 のページ中の

ネット投資家 4 割が損抱える

をクリックした場合, 図 5.14 が開かれる. このときリンクの参照位置となるのはページ左部にある

ネット投資家, ライブドア騒動で４割が“ 損 ”

であるが, アンカー文字列と参照位置文字列の間には言い換えが存在し, システムは正しくリンクの参照位置を同定することができない. しかしページの右部にトピックスの一覧を示したセグメントがあり, その中にアンカー文字列とまったく同じ文字列が含まれており, システムはこの文字列をリンクの参照位置として検出してしまう. 正しいリンクの参照位置が同定できなかった状態で誤同定した文字列をリンクの参照位置としてユーザに与えるならば, 無用な手間を生じさせる. このエラーを回避するためには, 言い換えの検出を行う

ことのできる手法を導入すると共に, セグメントの情報を利用してメニューの読み飛ばしを行ったり, ユーザに対して見出し等のセグメントの情報を与えるなどして対応する必要がある.

これらの結果より, 今後は単語ベクトルによる比較を用いた手法の開発を進めていく予定である. また,4 章で述べたページセグメンテーションの手法を利用し, そこで構築された構造を利用した目的位置の検出システムの開発も行っていきたいと考えている.

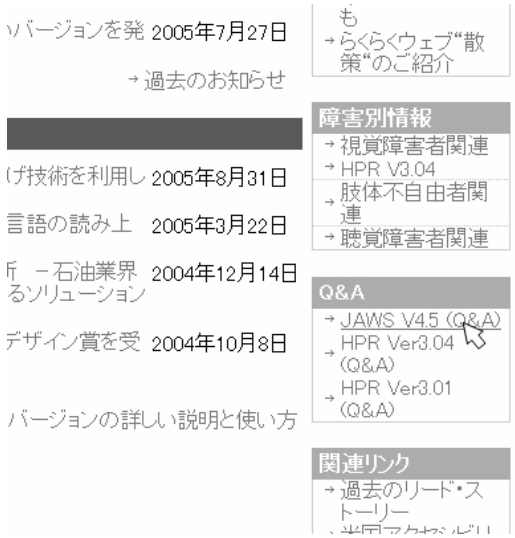


図 5.11: 言い換え例 3(リンク元ページ)


図 5.12: 言い換え例 3(リンク先ページ)

(IBM「バリアフリーの扉」ホームページより抜粋)

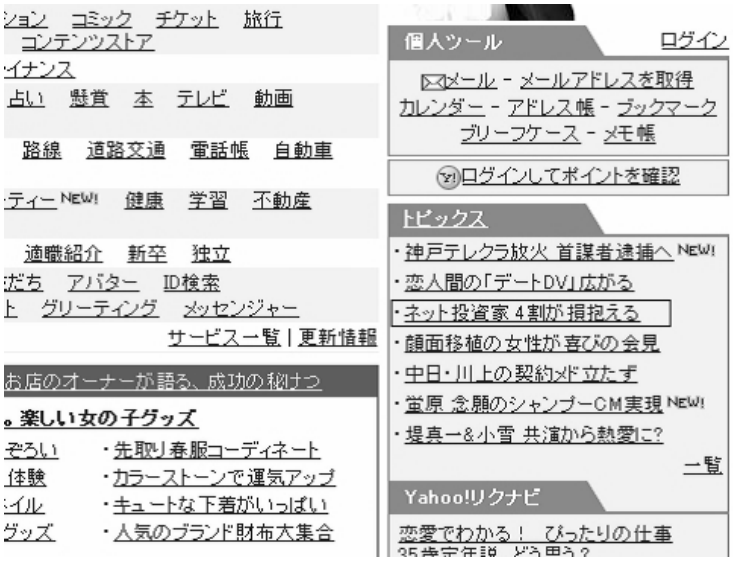


図 5.13: 誤同定例 (リンク元ページ)

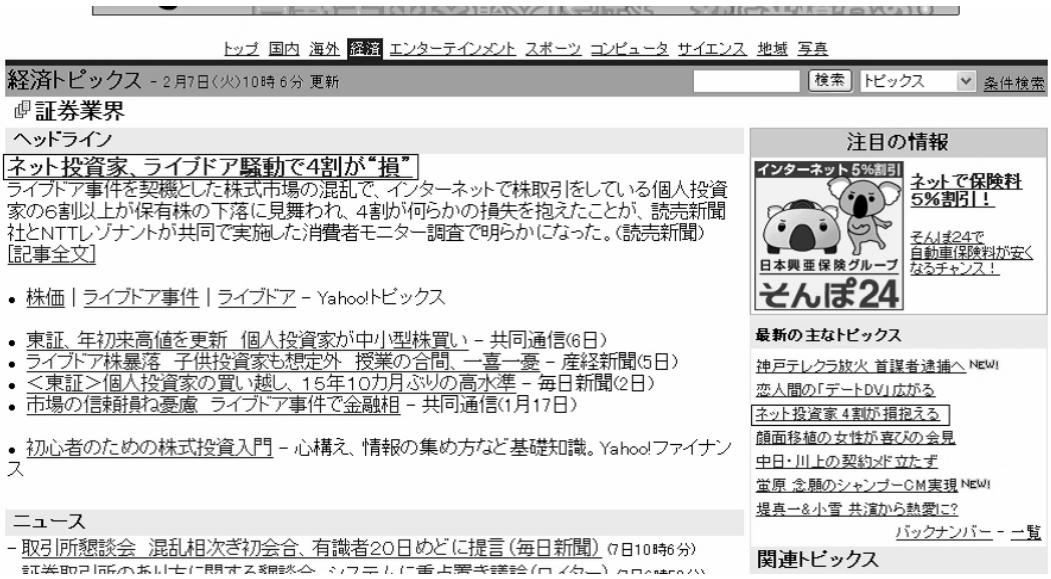


図 5.14: 誤同定例 (リンク先ページ)

(Yahoo!より抜粋)

# 第6章　おわりに

## 6.1　まとめ

本研究では, 視覚障害者がウェブブラウジングを行う際に用いる音声ブラウザの問題点を挙げ, その問題点を解決するために次の二つのシステムについて提案を行った.

1. ウェブページの構造を基にセグメントを検出し, それに従ってコンテンツの読み飛ばしをナビゲートする

2. リンクをクリックした際にユーザの目的位置が特定できる場合には, 開かれたページの読み上げをその箇所から開始する

それらのシステムを実装するため, ウェブページセグメンテーション手法とリンクの参照位置同定手法の開発を行い, それらの手法の有効性を示した.

ウェブページセグメンテーションにおいては,DOM を利用したセグメンテーションシステムを基礎とし, そこで検出されたセグメントに対し分割やマージといった新たな手法を適用してきた. それらの手法の有効性を確認するために, 各 20 個のウェブページを用いたクローズテストとオープンテストを行った. その結果, クローズテストでは正解率で 49%, 開始位置だけで言えば 64%のセグメントを検出することが, オープンテストでは正解率でそれぞれ 33%,58%のセグメントを検出することが確認された.

リンクの参照位置の同定においては, パターンマッチを用いたシステムを作成した. その際の手法として, アンカー文字列の長さとリンク先ページ内の文字列の長さを比較し, 短い方をパターンとしてマッチさせた. このような手法をとることによって, 完全一致では検出することのできない文字列を検出することが可能となっている. それらの手法の有効性を確認するために,20 種 100 ページに対する実験を行った. 実験の結果,100 ページ中 66 ページにおいてページ内位置を同定できることが確認された. さらにシステムの有効性を確認するため, 読み上げを省略可能なテキスト要素数の変化を算出すると, 本システムによって 70%の省略が実現可能であることが確認された. システムがリンクの参照位置を同定できた 66 ページだけで言えば, システムを利用しない場合の読み上げが必要なテキスト要素数の平均は 75.1 だったのに対し, システム使用時のテキスト要素数の平均は 2.2 であり, 約 97%のテキスト要素を省略できることが確認された. このことから確かに本システムが有効であるといえる.

## 6.2 今後の課題

本研究で作成したウェブページセグメンテーションシステムは, 実用に耐え得るものではない. これは実験の結果からも明らかである. 従ってさらに多くのウェブページを用いて実験を行い, エラーを解析し, 現在用いていない様々な情報を用いた手法を適用し, 精度の向上を図っていく必要がある. それらの情報として, 言語的特徴やレンダリングされた際に現れるレイアウト特徴といったものが挙げられる.

また現行のセグメンテーションシステムは最小の単位のセグメントを検出することを目的としている. 従って, 現在はセグメントの定義にある入れ子の状態を許していない. 今後は現在の手法の再帰的適用を含め, セグメントを含んだセグメントの検出を行う手法の開発を行っていく必要がある. 構造構築の手法のひとつの方法として, 本研究で開発したセグメントのマージシステムを改良することで対応できると考えられる.

さらに本研究の主要な目的は視覚障害者支援である. 従ってシステムが作成したセグメントをどのように音声ブラウザに適用するかといったことも考えていかなければならない. ユーザに有益な情報を与えるためにはセグメント情報をユーザにとって有益な情報に変換する必要がある. 例えば, ヘッダを持つセグメントに関してはどのようにヘッダを特定するかといったヘッダ検出タスクが, ヘッダを持たないセグメントに関してはその中に含まれるキーワード抽出タスクが必要となる.

またユーザインタラクションも今後の課題のひとつである. つまり, どのようにユーザに対してセグメントの情報を提示することが有益かといった問題である. ヘッダを持つセグメントに関しては, ヘッダを読み上げるだけで良いのか, 構造やキーワードといったものを伝えることができないだろうか, といったことも考えていかなければならない. ヘッダを持たないセグメントに関しては, どのような情報を提示すれば良いのか, さらに考察を進める必要があるが, ひとつの方法としてはキーワード抽出手法を適用することができるのではないかと考えられる. この問題に関しては, さらに実際のユーザの意見も取り入れていかなければならないため, 今後は視覚障害者の方々から意見を頂きつつ開発を進めていく予定である.

リンクの参照位置の同定を行うシステムに関しては, パターンマッチを利用した手法によってかなりのページで位置を同定することができた. しかし第 5 章の考察でも述べたように, アンカー文字列とリンク先の文字列の間に言い換えが行われている場合が存在し, そのような状況に対応するため, 単語ベクトルによる手法の導入を行う必要がある. 単語ベクトルによる比較を挙げた理由は, エラー解析を行った際にアンカー文字列とリンク先の文字列の間にはテキストで使用される単語までもが変化するような複雑な言い換えは存在しなかったためである. 従って今後は単語ベクトルを用いた手法を導入し, さらに多くのウェブページにおいて実験を重ね, 本システムの汎用性を高めると共に有効性を実証していきたいと考えている.

さらにこの目的位置同定システムにおいても, やはりユーザインタラクションは大きな課題のひとつとなる. 本システムを実際のソフトに実装した場合, システムが目的位置と思われるテキストを検出した時, ユーザに対してどのような形でその情報を提供するかと

いった問題である. それらの問題に対処するため, 今後は目的位置同定システムに対してページセグメンーションの情報を導入し, ヘッダやキーワードの抽出といったユーザインタラクションの際に有用な情報を取り出す手法の開発も行っていく必要がある. それらを含めシステムの実際の動作に関する部分においては, 実際のユーザの方からの意見を参考にシステムを作成していかなければならないであろう.

本研究の最終的な目標は, 視覚障害者のウェブブラウジングをより適切に支援することのできる音声ブラウザの開発である. 本研究で作成したシステムの性能を向上させ, 音声ブラウザに搭載することによって, ユーザが自分の知りたい情報に素早く辿り着くことのできる視覚障害者支援ツールを作成することができるであろう.

# 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, 白井清昭先生から様々な御指導を賜り, 心より感謝致します. また, 山田寛康助手, 中村誠助手, 同期の皆様には大変お世話になりました. 特に実験データの作成に携わって下さった中村達也氏には感謝しております.

さらに, 本研究において中山信雄様には視覚障害者の立場から様々な実際的な意見を頂きました.

お世話になった皆様に心から御礼申し上げます.

# 参考文献


[1] 前田潤治, 高木啓伸, 福田健太郎, 浅川智恵子, アノテーションに基づくウェブページのダイジェスト手法, 電子情報通信学会技術研究報告 (福祉情報工学), WIT2001-14, pp. 25-30, Aug.

[2] 浅川智恵子, 伊藤隆, 視覚障害者向け Web アクセスシステムにおける HTML タグの音声変換方式について, 自然言語処理シンポジウム, Nov. 1997.

[3] 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, 情報処理学会研究報告, 2003-NL-154, pp.185-192, 2003.

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

[5] 松本吉司, 乾健太郎, 松本裕治, Web ページのテキストセグメント階層構造の抽出, 言語処理学会第 11 回年次大会論文集, pp., March 2005.

[6] Yudong Yang and HongJiang Zhang, HTML Page Analysis Based on Visual Cues, Proceedings of the Sixth Internatinal Conference on Document Analysis and Recognition(ICDAR 2001).

[7] Extracting ontologies from World Wide Web via HTML tables, Minoru Yoshida, Kentaro Torisawa and Jun'ichi Tsujii, In Proceedings of Pacific Association for Computational Linguistics 2001 (PACLING 2001), Kitakyushu, Japan, September, 2001

[8] Tidy, http://tidy.sourceforge.net/.

[9] freqt, http://www.chasen.org/ taku/software/freqt/.

# 付 録A　実験結果(詳細)

## A.1　クローズテスト

オープンテストにおける各ウェブページの precision, recall, start position(recall), no cross(precision) をそれぞれ表に示す. 使用されたページの詳細は以下の表のとおりである.

表 A.1: ページ詳細

| page name | page title | URL |
|---|---|---|
| chunichi_0501.htm | 中日新聞ホームページ | http://www.chunichi.co.jp/ |
| gnavi_0501.htm | ぐるめ情報検索サイト ぐるなび | http://gnavi.joy.ne.jp/ |
| gov_01.htm | 電子政府の総合窓口 | http://www.e-gov.go.jp/ |
| ishikawa_0501.htm | 石川県ホームページ | http://www.pref.ishikawa.jp/ |
| library_01.htm | 北陸先端科学技術大学院大学 付属図書館 | http://www.jaist.ac.jp/library/index-jp.html |
| nomi_0501.htm | 能美市ホームページ | http://www.city.nomi.ishikawa.jp/ |
| ono_01.htm | Brilliant ◆ Shinji | http://www5.plala.or.jp/ichishin/subtop.htm |
| perl_01.htm | Perl for Newbie（Perl 初心者の部屋） | http://www.harukaze.net/ mishima/perl/index.html |
| perl_02.htm | beginners CGI　- 初歩から始める初心者向け Perl-CGI 講座 - | http://www.aimix.jp/cgi/ |
| search_0512.htm | Yahoo!検索 - 自然言語処理 | 省略 |
| shirai_01.htm | 研究者総覧 | http://www.jaist.ac.jp/ kkgi/thisyear/soj/00275soj.html |
| shop_01.htm | Yahoo!ショッピング - ウェブショップタテ | http://store.yahoo.co.jp/shop-tate/index.html |
| shop_02.htm | サッカーショップ ユニオンスポーツ | http://www.unionspo.com/ |
| shop_03.htm | イオンのオンラインショッピングサイト ～aeonshop.com～ | http://www.aeonshop.com/ |
| sportsnavi_0501.htm | スポーツナビ | http://sportsnavi.yahoo.co.jp/ |
| tel_01.htm | タウン検索なら i タウンページ | http://itp.ne.jp/ |
| unix_01.htm | UNIX 一年生 | http://www.tokaido.co.jp/syoko/index.html |
| we_01.htm | ZAPAnet ウイニングイレブン攻略情報 | http://zapanet.info/we/ |
| we_02.htm | ウイニングイレブン 9 必勝ガイド | http://winning-guide.com/ |
| yahoo_0501.htm | Yahoo! Japan | http://www.yahoo.co.jp/ |

表 A.2: ページ別結果 (SYS1)

| page name | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|
| chunichi_0501.htm | 8 | 24 | 33.33% | 27 | 29.63% | 48.15% | 100% |
| gnavi_0501.htm | 30 | 48 | 62.50% | 61 | 49.18% | 65.57% | 100% |
| gov_01.htm | 2 | 17 | 11.76% | 14 | 14.29% | 50.00% | 100% |
| ishikawa_0501.htm | 11 | 24 | 45.83% | 30 | 36.67% | 43.33% | 100% |
| library_01.htm | 3 | 11 | 27.27% | 13 | 23.08% | 30.77% | 100% |
| nomi_0501.htm | 4 | 9 | 44.44% | 15 | 26.67% | 46.67% | 100% |
| ono_01.htm | 2 | 6 | 33.33% | 11 | 18.18% | 27.27% | 100% |
| perl_01.htm | 33 | 55 | 60.00% | 48 | 68.75% | 91.67% | 100% |
| perl_02.htm | 6 | 9 | 66.67% | 12 | 50.00% | 66.67% | 100% |
| search_0512.htm | 3 | 21 | 14.29% | 25 | 12.00% | 40.00% | 100% |
| shirai_01.htm | 2 | 6 | 33.33% | 10 | 20.00% | 30.00% | 100% |
| shop_01.htm | 18 | 21 | 85.71% | 26 | 69.23% | 73.08% | 100% |
| shop_02.htm | 4 | 11 | 36.36% | 11 | 36.36% | 54.55% | 100% |
| shop_03.htm | 16 | 34 | 47.06% | 48 | 33.33% | 39.58% | 100% |
| sportsnavi_0501.htm | 4 | 22 | 18.18% | 25 | 16.00% | 36.00% | 100% |
| tel_01.htm | 4 | 28 | 14.29% | 19 | 21.05% | 68.42% | 100% |
| unix_01.htm | 0 | 8 | 0.00% | 5 | 0.00% | 0.00% | 100% |
| we_01.htm | 4 | 20 | 20.00% | 14 | 28.57% | 50.00% | 95.00% |
| we_02.htm | 1 | 33 | 3.03% | 25 | 4.00% | 52.00% | 93.94% |
| yahoo_0501.htm | 2 | 22 | 9.09% | 42 | 4.76% | 16.67% | 100% |
| **Average** | **7.9** | **21.5** | **33.32%** | **24.1** | **28.09%** | **46.52%** | **99.45%** |

表 A.3: タイプ別結果 (SYS1)

| type name | num | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|---|
| news | 2 | 6.0 | 23.0 | 25.76% | 26.0 | 22.82% | 42.08% | 100% |
| portal | 4 | 9.8 | 29.8 | 25.05% | 36.8 | 21.75% | 47.67% | 100% |
| business | 5 | 4.4 | 13.4 | 32.53% | 28.3 | 24.14% | 40.15% | 100% |
| shopping | 3 | 12.7 | 22.0 | 56.38% | 28.3 | 46.31% | 55.74% | 100% |
| personal | 6 | 7.7 | 21.8 | 30.51% | 19.2 | 28.25% | 47.94% | 98.16% |

表 A.4: ページ別結果 (SYS2)

| page name | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|
| chunichi_0501.htm | 11 | 27 | 40.74% | 27 | 40.74% | 51.85% | 88.89% |
| gnavi_0501.htm | 30 | 52 | 57.69% | 61 | 49.18% | 65.57% | 98.08% |
| gov_01.htm | 2 | 20 | 10.00% | 14 | 14.29% | 50.00% | 100% |
| ishikawa_0501.htm | 11 | 26 | 42.31% | 30 | 36.67% | 43.33% | 100% |
| library_01.htm | 3 | 15 | 20.00% | 13 | 23.08% | 30.77% | 100% |
| nomi_0501.htm | 4 | 9 | 44.44% | 15 | 26.67% | 46.67% | 100% |
| ono_01.htm | 2 | 6 | 33.33% | 11 | 18.18% | 27.27% | 100% |
| perl_01.htm | 33 | 55 | 60.00% | 48 | 68.75% | 91.67% | 100% |
| perl_02.htm | 6 | 9 | 66.67% | 12 | 50.00% | 66.67% | 100% |
| search_0512.htm | 13 | 30 | 43.33% | 25 | 52.00% | 76.00% | 100% |
| shirai_01.htm | 2 | 6 | 33.33% | 10 | 20.00% | 30.00% | 100% |
| shop_01.htm | 18 | 21 | 85.71% | 26 | 69.23% | 73.08% | 100% |
| shop_02.htm | 4 | 20 | 20.00% | 11 | 36.36% | 54.55% | 100% |
| shop_03.htm | 17 | 39 | 43.59% | 48 | 35.42% | 39.58% | 100% |
| sportsnavi_0501.htm | 4 | 28 | 14.29% | 25 | 16.00% | 36.00% | 100% |
| tel_01.htm | 3 | 33 | 9.09% | 19 | 15.79% | 68.42% | 100% |
| unix_01.htm | 0 | 8 | 0.00% | 5 | 0.00% | 0.00% | 100% |
| we_01.htm | 4 | 20 | 20.00% | 14 | 28.57% | 50.00% | 95.00% |
| we_02.htm | 1 | 33 | 3.03% | 25 | 4.00% | 52.00% | 93.94% |
| yahoo_0501.htm | 2 | 22 | 9.09% | 42 | 4.76% | 16.67% | 100% |
| **Average** | **8.5** | **24.0** | **32.83%** | **24.1** | **30.48%** | **48.51%** | **98.80%** |

表 A.5: タイプ別結果 (SYS2)

| type name | num | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|---|
| news | 2 | 7.5 | 27.5 | 27.52% | 26.0 | 28.37% | 43.93% | 94.45% |
| portal | 4 | 12.0 | 34.3 | 29.80% | 36.8 | 30.43% | 56.67% | 99.52% |
| business | 5 | 4.4 | 15.2 | 30.02% | 28.3 | 24.14% | 40.15% | 100% |
| shopping | 3 | 13.0 | 26.7 | 49.77% | 28.3 | 47.00% | 55.74% | 100% |
| personal | 6 | 7.7 | 21.8 | 30.51% | 19.2 | 28.25% | 47.94% | 98.16% |

表 A.6: ページ別結果 (SYS3)

| page name | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|
| chunichi_0501.htm | 11 | 24 | 45.83% | 27 | 40.74% | 44.44% | 95.83% |
| gnavi_0501.htm | 34 | 53 | 64.15% | 61 | 55.74% | 67.21% | 100% |
| gov_01.htm | 2 | 17 | 11.76% | 14 | 14.29% | 28.57% | 100% |
| ishikawa_0501.htm | 15 | 28 | 53.57% | 30 | 50.00% | 50.00% | 100% |
| library_01.htm | 3 | 20 | 15.00% | 13 | 23.08% | 38.46% | 100% |
| nomi_0501.htm | 10 | 13 | 76.92% | 15 | 66.67% | 73.33% | 100% |
| ono_01.htm | 8 | 19 | 42.11% | 11 | 72.73% | 72.73% | 100% |
| perl_01.htm | 33 | 55 | 60.00% | 48 | 68.75% | 91.67% | 100% |
| perl_02.htm | 5 | 16 | 31.25% | 12 | 41.67% | 50.00% | 100% |
| search_0512.htm | 14 | 30 | 46.67% | 25 | 56.00% | 76.00% | 100% |
| shirai_01.htm | 2 | 6 | 33.33% | 10 | 20.00% | 30.00% | 100% |
| shop_01.htm | 18 | 20 | 90.00% | 26 | 69.23% | 73.08% | 100% |
| shop_02.htm | 7 | 27 | 25.93% | 11 | 63.64% | 63.64% | 100% |
| shop_03.htm | 17 | 39 | 43.59% | 48 | 35.42% | 39.58% | 100% |
| sportsnavi_0501.htm | 3 | 44 | 6.82% | 25 | 12.00% | 32.00% | 100% |
| tel_01.htm | 2 | 42 | 4.76% | 19 | 10.53% | 68.42% | 100% |
| unix_01.htm | 0 | 8 | 0.00% | 5 | 0.00% | 0.00% | 100% |
| we_01.htm | 4 | 20 | 20.00% | 14 | 28.57% | 50.00% | 95.00% |
| we_02.htm | 1 | 29 | 3.45% | 25 | 4.00% | 36.00% | 93.10% |
| yahoo_0501.htm | 18 | 56 | 32.14% | 42 | 42.86% | 50.00% | 100% |
| **Average** | **10.4** | **28.3** | **35.36%** | **24.1** | **38.80%** | **51.76%** | **99.20%** |

表 A.7: タイプ別結果 (SYS3)

| type name | num | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|---|
| news | 2 | 7.0 | 34.0 | 26.33% | 26.0 | 26.37% | 38.22% | 97.92% |
| portal | 4 | 17.0 | 45.3 | 36.93% | 36.8 | 41.28% | 65.41% | 100% |
| business | 5 | 6.4 | 16.8 | 38.12% | 28.3 | 34.81% | 44.07% | 100% |
| shopping | 3 | 14.0 | 28.7 | 53.17% | 28.3 | 56.10% | 58.77% | 100% |
| personal | 6 | 8.5 | 24.5 | 26.14% | 19.2 | 35.95% | 50.07% | 98.02% |

表 A.8: ページ別結果 (SYS4)

| page name | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|
| chunichi_0501.htm | 14 | 24 | 58.33% | 27 | 51.85% | 59.26% | 91.67% |
| gnavi_0501.htm | 34 | 53 | 64.15% | 61 | 55.74% | 65.57% | 100% |
| gov_01.htm | 5 | 17 | 29.41% | 14 | 35.71% | 50.00% | 100% |
| ishikawa_0501.htm | 23 | 28 | 82.14% | 30 | 76.67% | 76.67% | 100% |
| library_01.htm | 5 | 20 | 25.00% | 13 | 38.46% | 76.92% | 100% |
| nomi_0501.htm | 8 | 13 | 61.54% | 15 | 53.33% | 60.00% | 100% |
| ono_01.htm | 8 | 19 | 42.11% | 11 | 72.73% | 90.91% | 100% |
| perl_01.htm | 33 | 55 | 60.00% | 48 | 68.75% | 93.75% | 100% |
| perl_02.htm | 7 | 16 | 43.75% | 12 | 58.33% | 75.00% | 100% |
| search_0512.htm | 13 | 30 | 43.33% | 25 | 52.00% | 72.00% | 100% |
| shirai_01.htm | 4 | 6 | 66.67% | 10 | 40.00% | 50.00% | 100% |
| shop_01.htm | 18 | 20 | 90.00% | 26 | 69.23% | 73.08% | 100% |
| shop_02.htm | 6 | 27 | 22.22% | 11 | 54.55% | 54.55% | 100% |
| shop_03.htm | 21 | 39 | 53.85% | 48 | 43.75% | 50.00% | 100% |
| sportsnavi_0501.htm | 3 | 44 | 6.82% | 25 | 12.00% | 32.00% | 100% |
| tel_01.htm | 4 | 42 | 9.52% | 19 | 21.05% | 78.95% | 100% |
| unix_01.htm | 5 | 8 | 62.50% | 5 | 100% | 100% | 100% |
| we_01.htm | 4 | 20 | 20.00% | 14 | 28.57% | 35.71% | 90.00% |
| we_02.htm | 3 | 29 | 10.34% | 25 | 12.00% | 36.00% | 89.66% |
| yahoo_0501.htm | 18 | 56 | 32.14% | 42 | 42.86% | 57.14% | 100% |
| **Average** | **11.8** | **28.3** | **44.19%** | **24.1** | **49.38%** | **64.38%** | **98.57%** |

表 A.9: タイプ別結果 (SYS4)

| type name | num | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|---|
| news | 2 | 8.5 | 34.0 | 32.58% | 26.0 | 31.93% | 45.63% | 95.84% |
| portal | 4 | 17.3 | 45.3 | 37.29% | 36.8 | 42.91% | 68.42% | 100% |
| business | 5 | 9.0 | 16.8 | 52.95% | 28.3 | 48.83% | 62.72% | 100% |
| shopping | 3 | 15.0 | 28.7 | 55.36% | 28.3 | 55.84% | 59.21% | 100% |
| personal | 6 | 10.0 | 24.5 | 39.78% | 19.2 | 56.73% | 71.90% | 96.61% |

## A.2 オープンテスト

オープンテストにおける各ウェブページの precision, recall, start position(recall), no cross(precision) をそれぞれ表に示す.

表 A.10: ページ詳細

| page name | page title | URL |
|---|---|---|
| all_about.htm | All About（オールアバウト） | http://allabout.co.jp/ |
| art_photo.htm | Art Photo Site | http://artphoto-site.com/ |
| auto_1202.htm | Yahoo!自動車 | http://autos.yahoo.co.jp/ |
| dic_01.htm | IT 用語辞典 e-Words | http://e-words.jp/ |
| kinen.htm | 禁煙成功の輪を広げよう | http://www.kinenseikou.net/ |
| snowboard.htm | Snowboard-senmon.com | http://www.snowboard-senmon.com/ |
| goo_com.htm | goo コンピューター パソコン 購入比較 スペック比較 | http://computers.goo.ne.jp/ |
| goo_gurume.htm | goo グルメ&料理 | http://gourmet.goo.ne.jp/ |
| kisyou_tyou.htm | 気象庁 Japan Meteorological Agency | http://www.jma.go.jp/jma/index.html |
| teraterm.htm | Tera Term Home Page | http://hp.vector.co.jp/authors/VA002416/ |
| buturi_key.htm | 物理のかぎしっぽ | http://www12.plala.or.jp/ksp/ |
| japan_java.htm | 財団法人インターネット協会 | http://www.iajapan.org/ |
| kurashi_web.htm | KURASHI Web | http://kurashi.hi-ho.ne.jp/ |
| write_program.htm | プログラムを書こう！ | http://www.asahi-net.or.jp/ yf8k-kbys/ |
| car_yomi.htm | @CARS : YOMIURI ONLINE（読売新聞） | http://www.yomiuri.co.jp/atcars/ |
| japan_texinfo.htm | 日本語 TeX 情報 | http://oku.edu.mie-u.ac.jp/ okumura/texfaq/ |
| manabi_net.htm | ニチイ学館 医療・福祉のまなびネット | http://www.e-nichii.net/ |
| cygwin.htm | Using Cygwin | http://sohda.net/cygwin/ |
| jr_1026.htm | JR おでかけネット | http://www.jr-odekake.net/ |
| yomi_online.htm | YOMIURI ONLINE(読売新聞) | http://www.yomiuri.co.jp/ |

表 A.11: ページ別結果 (SYS1)

| page name | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|
| all_about.htm | 10 | 31 | 32.26% | 33 | 30.30% | 57.58% | 100% |
| art_photo.htm | 26 | 49 | 53.06% | 47 | 55.32% | 82.98% | 100% |
| auto_1202.htm | 4 | 30 | 13.33% | 33 | 12.12% | 42.42% | 100% |
| dic_01.htm | 0 | 24 | 0.00% | 14 | 0.00% | 14.29% | 100% |
| kinen.htm | 3 | 17 | 17.65% | 29 | 10.34% | 24.14% | 100% |
| snowboard.htm | 3 | 8 | 37.50% | 14 | 21.43% | 35.71% | 87.50% |
| goo_com.htm | 5 | 17 | 29.41% | 22 | 22.73% | 40.91% | 100% |
| goo_gurume.htm | 12 | 22 | 54.55% | 32 | 37.50% | 53.12% | 100% |
| kisyou_tyou.htm | 0 | 26 | 0.00% | 24 | 0.00% | 29.17% | 100% |
| teraterm.htm | 2 | 11 | 18.18% | 14 | 14.29% | 28.57% | 100% |
| buturi_key.htm | 1 | 13 | 7.69% | 6 | 16.67% | 16.67% | 100% |
| japan_java.htm | 6 | 11 | 54.55% | 11 | 54.55% | 54.55% | 100% |
| kurashi_web.htm | 12 | 17 | 70.59% | 36 | 33.33% | 38.89% | 100% |
| write_program.htm | 5 | 9 | 55.56% | 7 | 71.43% | 85.71% | 100% |
| car_yomi.htm | 4 | 32 | 12.50% | 31 | 12.90% | 38.71% | 96.88% |
| japan_texinfo.htm | 3 | 10 | 30.00% | 7 | 42.86% | 42.86% | 100% |
| manabi_net.htm | 5 | 20 | 25.00% | 22 | 22.73% | 40.91% | 100% |
| cygwin.htm | 1 | 15 | 6.67% | 16 | 6.25% | 37.50% | 93.33% |
| jr_1026.htm | 5 | 19 | 26.32% | 25 | 20.00% | 28.00% | 100% |
| yomi_online.htm | 10 | 46 | 21.74% | 34 | 29.41% | 73.53% | 97.83% |
| **Average** | **5.9** | **21.4** | **28.33%** | **22.9** | **25.71%** | **43.31%** | **98.78%** |

表 A.12: タイプ別結果 (SYS1)

| type name | num | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|---|
| news | 1 | 10 | 46 | 21.74% | 34 | 29.41% | 73.53% | 97.83% |
| portal | 9 | 8.4 | 25.6 | 33.69% | 29.1 | 25.07% | 44.96% | 98.26% |
| bissines | 4 | 4 | 19 | 26.47% | 20.5 | 24.32% | 38.16% | 100% |
| personal | 6 | 2.5 | 12.5 | 22.63% | 13.2 | 26.97% | 39.24% | 98.89% |

表 A.13: ページ別結果 (SYS2)

| page name | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|
| all_about.htm | 10 | 31 | 32.26% | 33 | 30.30% | 57.58% | 100% |
| art_photo.htm | 26 | 49 | 53.06% | 47 | 55.32% | 82.98% | 100% |
| auto_1202.htm | 3 | 36 | 8.33% | 33 | 9.09% | 39.39% | 100% |
| dic_01.htm | 0 | 24 | 0.00% | 14 | 0.00% | 14.29% | 100% |
| kinen.htm | 14 | 33 | 42.42% | 29 | 48.28% | 58.62% | 96.97% |
| snowboard.htm | 3 | 8 | 37.50% | 14 | 21.43% | 35.71% | 87.50% |
| goo_com.htm | 5 | 17 | 29.41% | 22 | 22.73% | 40.91% | 100% |
| goo_gurume.htm | 12 | 22 | 54.55% | 32 | 37.50% | 53.12% | 100% |
| kisyou_tyou.htm | 0 | 26 | 0.00% | 24 | 0.00% | 29.17% | 100% |
| teraterm.htm | 2 | 11 | 18.18% | 14 | 14.29% | 28.57% | 100% |
| buturi_key.htm | 1 | 13 | 7.69% | 6 | 16.67% | 16.67% | 100% |
| japan_java.htm | 6 | 11 | 54.55% | 11 | 54.55% | 54.55% | 100% |
| kurashi_web.htm | 12 | 19 | 63.16% | 36 | 33.33% | 36.11% | 100% |
| write_program.htm | 5 | 9 | 55.56% | 7 | 71.43% | 85.71% | 100% |
| car_yomi.htm | 4 | 32 | 12.50% | 31 | 12.90% | 38.71% | 96.88% |
| japan_texinfo.htm | 3 | 10 | 30.00% | 7 | 42.86% | 42.86% | 100% |
| manabi_net.htm | 5 | 20 | 25.00% | 22 | 22.73% | 40.91% | 100% |
| cygwin.htm | 1 | 15 | 6.67% | 16 | 6.25% | 37.50% | 93.33% |
| jr_1026.htm | 5 | 20 | 25.00% | 25 | 20.00% | 24.00% | 90.00% |
| yomi_online.htm | 10 | 129 | 7.75% | 34 | 29.41% | 73.53% | 93.80% |
| **Average** | **6.4** | **26.8** | **28.18%** | **22.9** | **27.45%** | **44.54%** | **97.92%** |

表 A.14: タイプ別結果 (SYS2)

| type name | num | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|---|
| news | 1 | 10 | 129 | 7.75% | 34 | 29.41% | 73.53% | 93.80% |
| portal | 9 | 8.3 | 26.4 | 32.31% | 29.1 | 24.73% | 44.31% | 98.26% |
| bissines | 4 | 4 | 19.3 | 26.14% | 20.5 | 24.32% | 37.16% | 97.50% |
| personal | 6 | 4.3 | 15.2 | 26.75% | 13.2 | 33.30% | 44.99% | 98.38% |

表 A.15: ページ別結果 (SYS3)

| page name | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|
| all_about.htm | 7 | 57 | 12.28% | 33 | 21.21% | 57.58% | 100% |
| art_photo.htm | 20 | 55 | 36.36% | 47 | 42.55% | 78.72% | 100% |
| auto_1202.htm | 8 | 67 | 11.94% | 33 | 24.24% | 57.58% | 100% |
| dic_01.htm | 0 | 23 | 0.00% | 14 | 0.00% | 0.00% | 100% |
| kinen.htm | 14 | 32 | 43.75% | 29 | 48.28% | 58.62% | 100% |
| snowboard.htm | 3 | 8 | 37.50% | 14 | 21.43% | 35.71% | 87.50% |
| goo_com.htm | 9 | 19 | 47.37% | 22 | 40.91% | 50.00% | 100% |
| goo_gurume.htm | 19 | 33 | 57.58% | 32 | 59.38% | 68.75% | 100% |
| kisyou_tyou.htm | 2 | 27 | 7.41% | 24 | 8.33% | 33.33% | 100% |
| teraterm.htm | 2 | 11 | 18.18% | 14 | 14.29% | 28.57% | 100% |
| buturi_key.htm | 1 | 13 | 7.69% | 6 | 16.67% | 16.67% | 100% |
| japan_java.htm | 3 | 16 | 18.75% | 11 | 27.27% | 54.55% | 100% |
| kurashi_web.htm | 12 | 21 | 57.14% | 36 | 33.33% | 33.33% | 100% |
| write_program.htm | 4 | 15 | 26.67% | 7 | 57.14% | 85.71% | 100% |
| car_yomi.htm | 4 | 30 | 13.33% | 31 | 12.90% | 35.48% | 96.67% |
| japan_texinfo.htm | 3 | 10 | 30.00% | 7 | 42.86% | 42.86% | 100% |
| manabi_net.htm | 2 | 32 | 6.25% | 22 | 9.09% | 31.82% | 100% |
| cygwin.htm | 1 | 15 | 6.67% | 16 | 6.25% | 37.50% | 93.33% |
| jr_1026.htm | 5 | 19 | 26.32% | 25 | 20.00% | 20.00% | 89.47% |
| yomi_online.htm | 10 | 65 | 15.38% | 34 | 29.41% | 52.94% | 98.46% |
| **Average** | **6.5** | **28.4** | **24.03%** | **22.9** | **26.78%** | **43.99%** | **98.27%** |

表 A.16: タイプ別結果 (SYS3)

| type name | num | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|---|
| news | 1 | 10 | 65 | 15.38% | 34 | 29.41% | 52.94% | 98.46% |
| portal | 9 | 9.1 | 34.8 | 30.39% | 29.1 | 28.44% | 46.35% | 98.24% |
| bissines | 4 | 3 | 23.5 | 14.68% | 20.5 | 16.17% | 34.93% | 97.37% |
| personal | 6 | 4.2 | 16 | 22.16% | 13.2 | 30.92% | 44.99% | 98.89% |

表 A.17: ページ別結果 (SYS4)

| page name | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|
| all_about.htm | 8 | 57 | 14.04% | 33 | 24.24% | 72.73% | 98.25% |
| art_photo.htm | 20 | 55 | 36.36% | 47 | 42.55% | 76.60% | 98.18% |
| auto_1202.htm | 9 | 67 | 13.43% | 33 | 27.27% | 60.61% | 94.03% |
| dic_01.htm | 3 | 23 | 13.04% | 14 | 21.43% | 50.00% | 100% |
| kinen.htm | 16 | 32 | 50.00% | 29 | 55.17% | 62.07% | 96.88% |
| snowboard.htm | 4 | 8 | 50.00% | 14 | 28.57% | 42.86% | 87.50% |
| goo_com.htm | 9 | 19 | 47.37% | 22 | 40.91% | 59.09% | 100% |
| goo_gurume.htm | 20 | 33 | 60.61% | 32 | 62.50% | 75.00% | 100% |
| kisyou_tyou.htm | 9 | 27 | 33.33% | 24 | 37.50% | 70.83% | 100% |
| teraterm.htm | 2 | 11 | 18.18% | 14 | 14.29% | 28.57% | 100% |
| buturi_key.htm | 2 | 13 | 15.38% | 6 | 33.33% | 83.33% | 100% |
| japan_java.htm | 3 | 16 | 18.75% | 11 | 27.27% | 54.55% | 100% |
| kurashi_web.htm | 10 | 21 | 47.62% | 36 | 27.78% | 33.33% | 100% |
| write_program.htm | 4 | 15 | 26.67% | 7 | 57.14% | 85.71% | 100% |
| car_yomi.htm | 10 | 30 | 33.33% | 31 | 32.26% | 54.84% | 80.00% |
| japan_texinfo.htm | 4 | 10 | 40.00% | 7 | 57.14% | 85.71% | 100% |
| manabi_net.htm | 2 | 32 | 6.25% | 22 | 9.09% | 27.27% | 96.88% |
| cygwin.htm | 5 | 15 | 33.33% | 16 | 31.25% | 68.75% | 93.33% |
| jr_1026.htm | 5 | 19 | 26.32% | 25 | 20.00% | 20.00% | 89.47% |
| yomi_online.htm | 9 | 65 | 13.85% | 34 | 26.47% | 47.06% | 98.46% |
| **Average** | **7.7** | **28.4** | **29.89%** | **22.9** | **33.81%** | **57.95%** | **96.66%** |

表 A.18: タイプ別結果 (SYS4)

| type name | num | right | precision | | recall | | SP | cross |
|---|---|---|---|---|---|---|---|---|
| news | 1 | 9 | 65 | 13.85% | 34 | 26.47% | 47.06% | 98.46% |
| portal | 9 | 10.3 | 34.8 | 35.09% | 29.1 | 34.17% | 58.34% | 95.33% |
| bissines | 4 | 4.8 | 23.5 | 21.16% | 20.5 | 23.47% | 43.16% | 96.59% |
| personal | 6 | 5.5 | 16 | 30.59% | 13.2 | 41.39% | 69.02% | 98.37% |