JAIST Repository
https://dspace.jaist.ac.jp/

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

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

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

加藤 邦彦 (410030)

北陸先端科学技術大学院大学 情報科学研究科

2006 年 2 月 9 日

キーワード: 音声ブラウザ, ウェブページ構造解析, セグメント,DOM, リンクのアンカー文字列.

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. 視覚障害者はウェブ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. それらのツールには, 画面の情報を音声化するスクリーンリーダーやウェブページに特化した機能を持つ音声ブラウザなどがある. しかしそれら既存のツールには, 知りたい情報が書かれている場所に辿り着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.

この問題に対処するため, 本研究で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以下の二つの方法を提案する.(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ト検出を行い, 不要な情報の読み飛ばしをナビゲーションする.(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. これらの機能を持った音声ブラウザを作成することによって, 音声のみでも快適にウェブブラウジングを行うことができると期待される.

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システムを実現するため, ウェブページセグメント検出手法の開発を行った. 提案手法は DOM を利用し, ウェブページの中で構造として使用されるタグを基にセグメントを検出する. さらに検出されたセグメントに対し, イメージを用いたセグメント分割手法やテーブル内の DOM の部分木を利用したセグメント分割手法, セグメントマージ手法を適用し, セグメント検出の正確性を向上させている. イメージを用いた分割とは, セグメント中に存在する最多のイメージをセグメントの境界とみなす手法である. テーブル内の DOM の部分木を利用した分割とは, テーブル内の`<td>`,`<tr>`タグを根とした DOM の部分木を取り出し, それらの類似度が高ければ各部分木をひとつのセグメントとして分割する手法である. また, セグメントマージ手法とは, ヘッダ部分を含まないセグメントのためにヘッダ部分を検出し, セグメントとマージする手法である.

Copyright © 2006 by Katou Kunihiko

提案手法の有効性を確認するため, 実在するウェブページを用いて実験を行った. まず 20 個のページを収集し, 人手で正解データを作成した. それらのページを分析し, システムの開発を行い, テストした結果,49%のセグメントを検出することができた. さらに本システムの目的が不要テキストの読み飛ばしであり, セグメントの開始位置を正確に検出することができればある程度有効な読み飛ばしを行えると考え, 開始位置のみの再現率を求めた結果,64%となった. また, 検出セグメントが複数の正解セグメントをまたいでいるものの割合は 1.5%未満であった. これらの結果より, 提案手法がセグメント検出に有効であると言える. また,DOM に基づく手法にセグメント分割手法やマージ手法を適用することによる精度の向上も確認された.

同システムを用いて新たな 20 ページに対してテストを行った結果,34%の正解率を得た. 開始位置のみの再現率は 58%であり, 検出セグメントのエラー率は最大でも 4%未満であった. 今後も手法の改良によってより正確なセグメント検出が実現できると期待される.

リンクを辿った際にユーザの目的位置からページの読み上げを開始するシステムに関しては, ユーザの目的位置となるリンクの参照位置を同定する手法の開発を行った. 提案手法はリンクのアンカー文字列とリンク先ページ内のテキストの間でパターンマッチを行い, マッチしたものを参照位置として検出するものである. その際, アンカー文字列とページ内テキストのテキスト長を比較し, 短い側の文字列をパターンとして長い側のテキストにマッチさせる方法を取っている. この方法によって, アンカー文字列やページ内テキストのどちらか片方に余分な文字列が付加されたとしても, ページ内の文字列を検出できる.

実験では実際のウェブページ 20 ページから各 5 個計 100 個のリンクを選び, それらに対し提案手法を適用した. その結果,67 ページにおいて正しくリンクの参照位置を検出することができた. 同定できた 67 ページにおいて, ページの読み上げの際に目的位置に辿り着くまでに読み上げられるテキスト数の平均を求めた結果, システム未使用時が 76.1 だったのに対し, システム使用時には 1.5 となり, 提案システムがユーザの手間を省略することが実証された.

以上の結果から, 本研究で提案したシステムが, さらなる手法の改良によって性能を向上させる必要はあるものの, 視覚障害者のウェブブラウジングを適切に支援することが確認された.