K-4 文書サムネイルによる視覚障碍者向けWeb閲覧方式の提案
Web Browsing Method with Document Thumbnail for Visually Impaired People

山口 俊光† 納富 一宏† 平松 明希子† 斉藤 恵一‡ 石井 博章§
Toshimitsu Yamaguchi Kazuhiro Notomi Akiko Hiramatsu Keichi Saito Hiroaki Ishii

1. はじめに

視覚障碍や聴覚障碍をはじめとする感覚障碍は外界からの情報の受容手段が制限されているため，情報の入手が困難である．人間が得ている外界情報のうち約80%は視覚によるといわれているほど，視覚に依存した情報受容を人間は行っている．この視覚による情報の受容が困難である視覚障碍者のための情報機器インタフェースの研究・開発は非常に重要なことである．

視覚障碍者のWeb閲覧環境としては音声ブラウザを利用した方法が一般的である．この方法ではHTML内に含まれるテキスト情報を1次元的に読み上げていく．この方法により，視覚障碍者の情報取得のための環境は大きく改善され，リアルタイムに情報を入手することが可能となった．しかしながら，既存の音声ブラウザを使用した方法では，晴眼者が通常，Webを閲覧する際に行っているような「斜め読み」や「拾い読み」は極めて困難，または不可能である．

また，PDAや携帯電話のような小型ディスプレイしか持たない情報機器を利用する晴眼者の場合でも，文書全体が見渡せないため，視覚障碍者と同様に「斜め読み」や「拾い読み」に相当するような行為を行うことは難しい．

本研究では，音声出力インタフェースを用いて「斜め読み」に対応したインタフェースの提供が可能であるかを検討する．

2. 提案システム

サーチエンジンなどで検索してきたWebページを閲覧する際，本文を最初から読み始める前に，文書全体を見渡し所々で拾い読みをしながら，実際に自分が求めている情報が記述されているかどうかを，簡単に確認することがよくある．音声による，1次元的な音声出力のみを頼りにWeb閲覧を行っているユーザの場合，このような，閲覧方法を行うことはきわめて難しい．

提案手法によるシステムの構成をFig.1に示す．HTML文書中からキーワードとなりうる語を複数抽出する．この語群を文書の特徴を表現しているものとして，「文書サムネイル」と呼ぶ．この文書サムネイルをスピーチエンジンにより音声化することで，ユーザに情報を提示する．

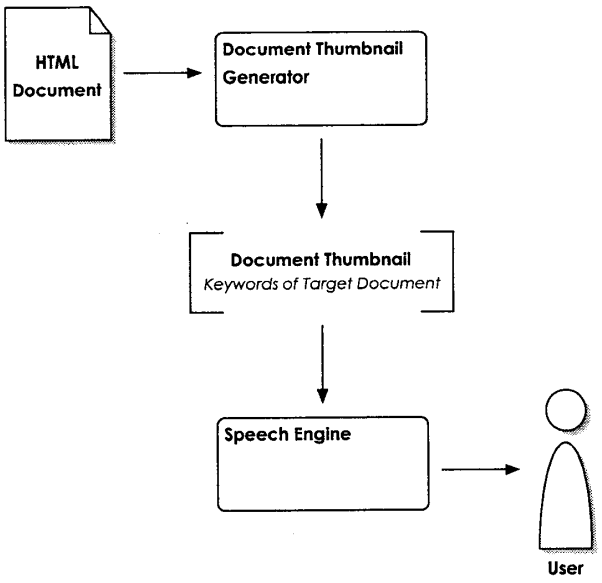

Fig. 1: System Structure

2.1 文書サムネイルの生成

文書サムネイルを生成するエンジンの構成をFig.2に示す．

まず，入力されたHTMLドキュメントから，HTMLによって意味付けされ「見出し」とされている部分を取り出す．HTMLはブラウザで閲覧した際の見た目を調節する役割のほかに，文書の論理的な構造を表現している．<H1>等の見出しタグで表現される文字列は文書の特徴を表している可能性が高い．この方法によりドキュメントの特徴を抽出しユーザに提示する手法はすでに提案され，有効性が示唆されている[1].

次に，HTML文書から本文に相当する部分を取り出し，形態素解析を行うことで，キーワードを抽出する．文書から複数のキーワードとなる可能性のある語を抽出する際．キーワードとなり得る語は文書中の自立語成分に含まれると考える．この際，見出しとして取り出しておいた文も併せて形態素解析を行い，自立語成分を抽出する．

形態素解析により得られた自立語成分の中からキーワードとなり得る語を抽出するために，自立語成分を評価していく必要がある．その評価方法としては以下のようなものが考えられる．

†神奈川工科大学情報工学科
‡東京電機大学超電導応用研究所
§神奈川工科大学福祉システム工学科

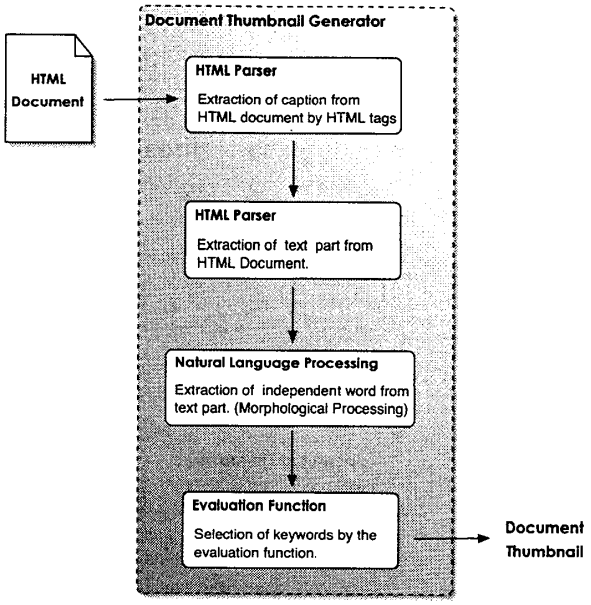

Fig. 2: Document Thumbnail Generator

- レイアウトから得られる情報
- 文字の修飾
- キーワード同士の関連
- 出現頻度

レイアウトからは，タグにより意味づけされていない見出し等を見つけだすことができると考える．例えば，前後に空行を含む独立した1行は見出しやタイトルなどで，強調したいと作成者が考えた文であると考えられる．ゆえに，ここに含まれている自立語成分はキーワードとなる可能性が高い．

文字の修飾はHTMLのタグによってなされる．文書中の他の部分と違う色を用いた部分や，違うフォントを用いた部分，アンダーラインがある部分などは，作成者が，強調したいと考えた部分であると考えられ，これもまたキーワードになる可能性が高い．

抽出されたキーワードの文中における出現位置を基に，キーワード間の関連性を推測することができると考える．関連性が高いキーワードは文中でも比較的近い位置にまとまっている可能性が高い．音声提示の順序等を決定する際キーワード間の関連性を考慮に入れることで，ユーザが文書の概要を把握しやすくなると考える．

出現頻度に関しては，繰り返し用いられる語も作成者が強調したいと考えている可能性が高い．しかし，繰り返される自立語はごく一般的な表現である可能性も高く，必ずしもキーワードになる可能性が高いとは言い切れない．カタカナ語，日本語文中における英単語，といった特殊な場合にはこの出現頻度は有効な評価手法であると考えられる．

ここまでに挙げた，評価手法により抽出されてきた自立語を評価していく．評価項目に当てはまる要素が多いほど，評価値が高くなるよう評価関数を定義し，評価値が高いものからキーワードとして選び出していく．

3. 評価手法

3.1 キーワードの評価

タイトルに用いられる自立語は，本文の内容を代表する語であると考え，以下のような評価手法を用いて提案手法の評価を行う．すでにタイトルが付けられているHTML文書をニュースサイト等のWebページから選び出し，本文から実際に抽出したキーワードとタイトル内に含まれる自立語成分とを比較し，抽出されたキーワードが適切なものであるかを評価する．

4. まとめ

文書サムネイルによる視覚障碍者向けWeb閲覧方式の提案について述べた．

今回は手法の提案のみにとどまったが，今後はユーザビリティの立場から実際に当事者の方に利用していただいて評価実験を行っていきたいと考えている．

参考文献

[1] 日本電信電話株式会社，“ネットワーク分散協調技術”，高齢者・障害者のための機能代行・支援通信システム技術の研究開発 平成11年度報告書，http://www.ucn21.com/h11/h11_5_1/h11_5_1.pd 2000