文書サムネイルを用いた視覚障碍者向けWeb閲覧方式の提案

山口 俊光[1]　納富 一宏[2]　平松 明希子[3]　斎藤 恵一[4]　石井 博章[5]

[1] 神奈川工科大学 博士前期課程 情報工学専攻
[2] 神奈川工科大学情報工学科
[3] 神奈川工科大学情報工学科 学部3年
[4] 東京電機大学超伝導応用研究所
[5] 神奈川工科大学福祉システム工学科

Proposal of Web Browsing Method with Document Thumbnail for Visually Impaired People

Toshimitsu YAMAGUCHI[1]　Kazuhiro NOTOMI[2]　Akiko HIRAMATSU[3]
Keiichi SAITO[4]　Hiroaki ISHII[5]

Abstract

In this article, we propose a new document browsing method for visually impaired people. It became possible to use computer and network cheaply. And visually impaired people's information acquisition environment is improving. However, It is hard to say that the optimal environment came to be provided.

We developed test system implemented in C++ on the Internet server to make Document Thumbnail. Our method is avaiable to read large documents with a speech engine.

Keywords: Visually Impaired People, Web Browsing, Natural Language Processing, Morphological Analysis, World Wide Web

1　はしがき

今日，コンピュータ本体やコンピュータネットワークへの接続料金が安価になり，コンピュータを利用した情報共有は一般的なものになりつつある．インターネットにおける最もユーザの多いアプリケーションはWorld Wide Web(Web)である．Webの急速な普及により，多くのユーザがインターネットを利用するようになった．インターネットは研究目的や軍事目的を中心にしていた時代を経て，電話や水道，ガスといった社会インフラの一部となりつつある．コンピュータ(特にパーソナルコンピュータ)が入手しやすくなったことで，視覚障碍者をはじめとする障碍者の情報受信，発信の環境は次第に改善されつつある．しかしながら，その環境は障碍者にとって最適なものとは言い難い．

視覚障碍は白内障等の疾患により，高齢化に伴い誰もが潜在的に直面する可能性のある障碍である．視覚障碍者数は平成14年4月に提出された「身体障害児・者実態調査」[1]によると全国で30万1千人となっているが，障害者手帳の発給を受けていない，病気や加齢に伴い視力が低下している人も含めると，実際はもっと多い数になることが考えられる．視覚障碍者の情報メディアとしては点字や音声が一般によく知られているが，中途失明者が点字を習得することは困難であり，点字による情報提供のみでは中途失明者に対して，適切に情報伝達が行えない．よって，音声による情報提示の重要性は今後ますます高まると考える．

健常者は感覚器情報のうち約80%を視覚から得

ていると言われているように，視覚障碍者の情報の受容は困難である．コンピュータ等の電子情報機器を使用する際も同様で．近年はテキスト以外の視覚による情報提示が多用されているため，その使用には訓練が必要になる．

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広く用いられている．ピンディスプレイを用いた点字による環境も提供されているが，音声ブラウザの方が特殊なデバイスを必要とせず，安価であるという特徴がある．音声ブラウザでは 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．もともと 2 次元にレイアウトされたテキスト情報を音声にメディア変換してしまうことで，情報は時間軸に沿って 1 次元化されてしまう．これにより，情報に対しシーケンシャルなアクセスしかできず，情報を俯瞰し概要を短時間に把握するのことは困難になる．多くの Web ページの中から自分の欲する情報を見つけだし，適切な情報を得るためには非常に多くの時間を必要としてしまう．晴眼者が情報の取捨選択を行う際には，いわゆる「斜め読み」,「拾い読み」を行い情報の概要を把握していくことで判断する場合が多い．既存の音声ブラウザを使用した方法では，その手法を用いて情報の概要を把握するのは極めて困難，または不可能である．また，PDA(Personal Digital/Data Assistants) 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．

本稿では，音声出力インタフェースを用いて「斜め読み」に相当するような情報を俯瞰する手法について提案する．

2 システム概要

2.1 文書サムネイル Document Thumbnail

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得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．

サムネイルとは画像を取り扱う際によく用いられる表示形態で，元画像を縮小し，アイコンとして画面に表示することで，ユーザに多くの画像の中から自分が欲している画像を容易に見つけだすことができるインタフェースを提供している．極端に縮小されているため，元の画像の詳細までは判別できないが，画像の概要を把握することは可能である．(Fig.2 参照)

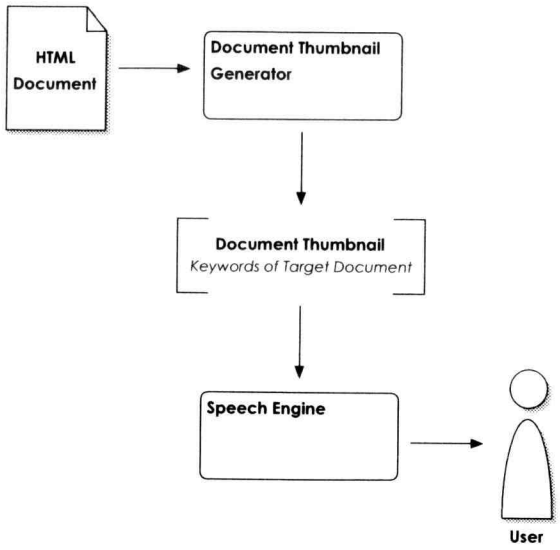

Fig 1: System Structure


Picture

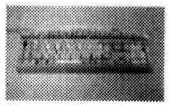
Picture Thumbnail

Fig 2: Picture Thumbnail

本稿では文書ファイルを閲覧する際にも全文を読む前に画像のサムネイルのような形で，文書の概要をユーザに対して提供する文書閲覧方法を提案する．文書から自然言語処理的手法を用いて文

書の特徴を抽出しユーザに対し提供する．この，提供するもののことを画像のそれになぞらえて，文書サムネイルと呼ぶことにする．

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．

2.2 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.3 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 [2]．

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．

- 出現頻度
- 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．

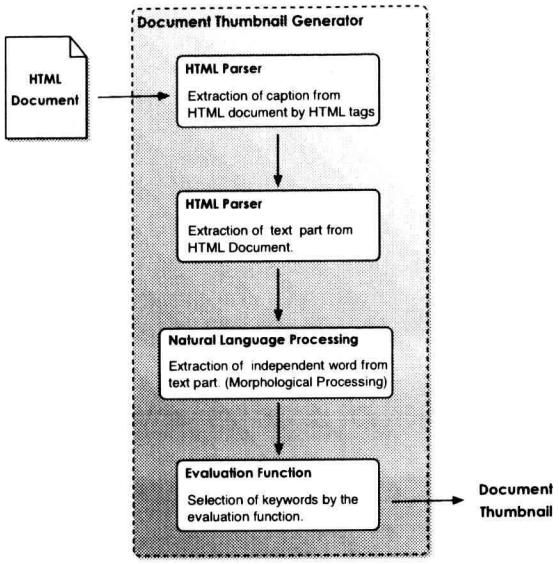

Fig 3: Document Thumbnail Generator

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．

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.

2.3 文書サムネイル生成システムの実装

文書サムネイルを生成するための実験的なシステムを実装した. 今回の実装は先に挙げた自立語成文の評価手法のうち, 出現頻度によるキーワード抽出のみに着目して行うこととした.

システムが完成し, 実際にサービスを提供する際にはユーザがWebブラウザが動作する環境だけを用意しておけばよいCGIを用いた方式が有効であると考える. そこで, 実用化を見据え本システムはLinux OS上で動作するWebサーバのCGIとして実装されている.

また, システム構築には動作が高速で, 一般に広く用いられてる言語であるC++を用いた. システムが実際に動作している様子をFig.4に示す. なお, このシステムは得られる文書サムネイルを確認するために構築したもので, 実際に視覚障碍者に提供するインタフェースではない.

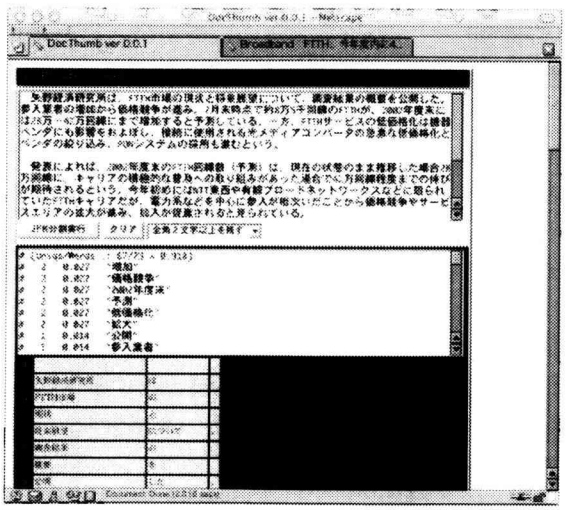

Fig 4: Screen Shot

画面最上部のフォームに自然言語で入力することで, 文書サムネイルを得ることができる.

このシステムではまず簡易な形態素解析手法としてJFK分解[5]を行い, 自立語成分を抽出する. 次に抽出された自立語成分の中から文書の特徴を表す語を選び出す. 今回の実装では, 抽出された自立語の出現頻度を元にランク付けを行い, 出現頻度が高いものほど, 文書の特徴を表している自立語として文書サムネイルとする. また, JFK分解は辞書を用いず文字種を元に形態素解析を行う

小泉純一郎首相は２３日夜（日本時間２４日未明）、コペンハーゲン市内のホテルで記者団に対し、内閣改造と自民党役員人事について、同党役員の任期が切れる３０日にも実施する意向を明らかにした。首相は民間人続投に与党内に異論があることに関して「大幅改造を期待している人の議論でしょ」と不快感を示...

— 中略 —

... さらに首相は公明、保守両党の意向を踏まえ、派閥推薦は受け付けず内閣改造を実施する考えを改めて強調した。その一方で、参院の閣僚については「衆院と参院は（事情が）ちょっと違う」と述べ、参院側の要望に配慮する姿勢を示した。 内閣改造、党役員人事をめぐっては、首相はすでに、福田長官、山崎幹事長らの続投を固めている。

Fig 5: Example Document1

手法なので,「漢字１文字＋送りがな」という場合,「漢字１文字」のみが自立語として取り出されてしまう. 漢字１文字では, 意味を把握しかねる場合があると考えるので, 自立語として扱う最低文字数を可変とし, １文字から５文字までの範囲で設定可能とした.

3 評価

実際のニュース記事を用いて, 動作実験を行った. 動作実験に用いたニュース記事は例文１が1019字, 例文２が728字である.

Fig.5に例文１を, Table.1に例文１から得られた文書サムネイルを示す. この文書サムネイルでは「漢字１文字+送りがな」という構成要素が漢字１文字だけで自立語のキーワードとしてあげられている.

Fig.6に例文２を, Table.2に例文２から得られた文書サムネイルを示す. こちらの文書サムネイル生成では, 自立語として形態素解析から得られた語のうち, ２文字以上のものだけを文書サムネイルを構成するキーワードとしているので, 例文１の文書サムネイルに含まれていたような漢字１文字だけのキーワードは抽出されない. 文書の特徴がより把握しやすい文書サムネイルになってい

Table 1: Keyword Extraction Result — Example Document 1

出現頻度	出現率	キーワード
6	0.046	"首相"
5	0.038	"内閣改造"
5	0.038	"考"
3	0.023	"述"
3	0.023	"臨時国会"
2	0.015	"関"
2	0.015	"示"
2	0.015	"意向"
2	0.015	"含"
2	0.015	"実施"
2	0.015	"衆院"
2	0.015	"会談"

ると思われる．

4 今後の展望

本稿では文書サムネイルの実装を行い提案するにとどまったが，今後は，実際にスピーチエンジンで文書サムネイルを読み上げる部分までの実装を行う．さらに，実際の視覚障碍者を被験者にした評価実験を行いキーワードのランク付けを行う関数等に反映させていく予定である．

筆者らは，これまでに自己組織化マップ (Self-Organizing Maps:SOM) を用いた臨床症例文書の分類手法を提案してきた [3][4]．SOM は教師無し学習を行う競合学習型のニューラルネットワークである．データ間の特徴類似度のよる汎用的なクラスタリング能力を持っている．これまでに行った提案では典型的な症例文書を用いて消化器系，循環器系，呼吸器系の 3 系をあらかじめ SOM に学習させておき，未知の症例がどの系の症例に似ているかを提案するシステムを構築した．このとき用いた手法を本研究に適用し，これから読み上げさせようとしている文書がどのようなジャンルに属しているかをコンピュータが提案するシステムが構築可能であると考えている．

5 むすび

文書サムネイルを用いた視覚障碍者向け Web 閲覧方式について述べた．特に今回は自然言語から自立語を取り出し，その出現頻度からキーワードを選び出し，文書サムネイルとしてテキスト出力する部分の実装を行った．今後は先に述べた展望に基づきさらに研究・実装を進めていく予定である．

東京電力は２４日、シュラウド（炉心隔壁）のひび割れ兆候を隠したとされる福島第１原発４号機と同第２原発２号機で、安全性を確認する自主検査を開始した。シュラウドのひび割れが指摘されている東電の３原発５基の中では最初の検査開始。原子力安全・保安院が電気事業法と原子炉等規制法に基づき、県と地元自治体が安全協定に基づき、それぞれ立ち入る中で実施する異例の自主検査となる。...

— 中略 —

... ４号機では、ＧＥ（ゼネラル・エレクトリック）社がシュラウド内側２カ所で三つのひび割れの兆候を発見したが、東電が報告書を書き換えるなどして国に報告せず、そのまま使用されている。これまでに東電、保安院ともに「安全性には問題ない」との判断を示しているが、ひび割れが確認されれば、運転再開には時間がかかりそうだ。

Fig 6: Example Document 2

参考文献

[1] 厚生労働省社会・援護局障害保健福祉部, "身体障害児・者実態調査", 平成 14 年 4 月, http://www.mhlw.go.jp/houdou/2002/04/h0411-2.html, 2002. 4

[2] 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.pdf, 2000

[3] 納富, 岡本, 山口, 他, "WWW による臨床症例検索システムの開発 — 自然言語処理と自己組織化マップを用いた疾患系分類 —", 第 61 回情報処理学会全国大会 4R-6, 2000. 10

Table 2: Keyword Extraction Result — Example Document 2

出現頻度	出現率	キーワード
4	0.057	"東電"
3	0.043	"確認"
3	0.043	"安全性"
3	0.043	"保安院"
2	0.029	"自主検査"
2	0.029	"指摘"
2	0.029	"シュラウド"
2	0.029	"兆候"
1	0.014	"炉心隔壁"
1	0.014	"２４日"
1	0.014	"開始"
1	0.014	"福島第１原発４号機"

[4] 納富, 山口, 他, "臨床症例データベース管理システムの構築 — 自己組織化マップによる情報の視覚化と検索インタフェース —", 第62回情報処理学会全国大会 6Q-1, 2001. 3

[5] 納富, 石井, "日本語文書構成支援システムにおける高速統語解析手法", 神奈川工科大学研究報告 B 理工学編 第20号, 1996. 3