I_047

文書構造要約化による情報提供システム

Information Communication System based on Document Structure Analysis and Summarization

関 峰伸† 永崎 健† 丸川 勝美†
Minenobu Seki Takeshi Nagasaki Katsumi Marukawa

1. 緒言

WWW 上には，膨大なコンテンツが存在するため,多く人は必要な情報の取捨選択に多大な時間を費やし，満足な結果を得られない場合がある。一般的に，ユーザは必要とする情報を探すために，大量に表示された検索結果の中から，1 ページずつ自分の目的にあうものを閲覧して精査する必要がある。この際，各ページ内には広告なども含めて様々な情報が含まれており，記述量が多くなる場合もある。そのため，ユーザが目的の記載部分をすぐに特定することは容易ではない。また，目的の情報を持つページが上位の検索結果にあがらないと，閲覧しなければならないページ数が多くなる。

本報告では，文書構造要約化を用いて，ユーザが目的の内容に効率的に到達できる情報提供システムを提案する。また，Adobe(R) Acrobat(R)を用いた PDF 文書構造要約プロトタイプを作成した。

2. 文書構造要約化による情報提供

2.1. コンセプト

提案する情報提供システムには 3 つの特徴がある。第 1 は，構造要約表示である。これは，広告，本文，タイトルなどの文書構成要素を区別し，章立てや構成要素間の関係付けを行うことで文書を構造化する。そして，その文書全体の構造を保持したまま，ユーザが着目する部分をハイライトし，関連のない部分を省略して表示することである。第 2 は，サイト毎に情報提供者が用意した関連ページ情報を提供することである。第 3 は，様々なデータ形式の文書を XML 化し管理することで統一的に扱うことである。これらの特徴により，効率的に有益な情報に到達可能になる。

本システムの例として，検索クエリー“OCR”を入力し，選択されたページを構造要約した例を図 1 に示す。文書内容や検索クエリーから重要な構成要素のみを表示することで，一見して内容を理解できるように表示される。すなわち，タイトルなどの全体構造を示す部分は残しつつ，検索クエリー“OCR”とは関係ないテキストや広告，ロゴなどが省略されて表示される。省略部分は，階層化されたデータとして保持されており，画面上でクリックすることでその内容を簡単に閲覧することができる。また，関連情報の提供についてはサイトを運用する情報提供者が「“OCR”に関心のあるユーザに“文書管理”のページを提示することが有益」と考え，“OCR”に対し“文書管理”を対応付けたキーワード関連表を事前に作成しておく。システムは検索時に関連表を引き，通常では検索結果に上がらない“文書管理”に関するページのリンク情報を提示する。

† （株）日立製作所中央研究所知能システム研究部

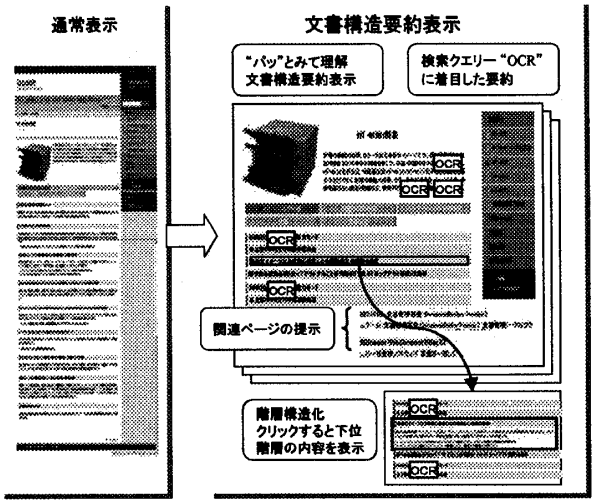

図 1 文書構造要約表示の例

2.2. 関連研究

本研究の目的に最も近い研究に，農工大品川等の WWW 閲覧支援の研究[1]がある。これは，HTML のタグ情報を解析することで文書を構造化し，キーワードを含む記述部分を強調して表示する。この解析手法は，論理構造を表す<H>タグを用いて解析しているが，<H>タグだけではデザインの融通性がきかないため,現実に存在する HTML 文書の多くは<H>タグを使用していない。したがって，この手法を実利用することは難しい。HTML を構造化するその他の手法として，タグの繰り返し構造を用いる手法[2]を始め，いくつかの手法が提案されているが限定された構造化にとどまっている。また，電子文書だけでなく,紙文書をスキャンした文書画像のレイアウト解析技術もある[3]。しかし，WWW 上のページのような複雑かつ多様な構造を持つ文書を汎用的に解析するのは難しい。

3. 文書構造要約化による情報提供システム

3.1. システム構成

提案システムの構成と処理フローを図 2 に示す。予め，情報提供者のサーバには，管理するページを文書構造化処理した文書構造化データ，検索クエリーに対応するキーワード表，インデックスファイルを作成しておく。

ユーザは検索クエリーを入力し，得られた検索結果からページを選択する。これに対し，情報提供サーバでは選択されたページの関連ページ付与と構造要約処理が行われ，構造要約ページが生成される。本システムの主な処理である文書構造化処理，構造要約処理を 3.2，3.3 で説明する。

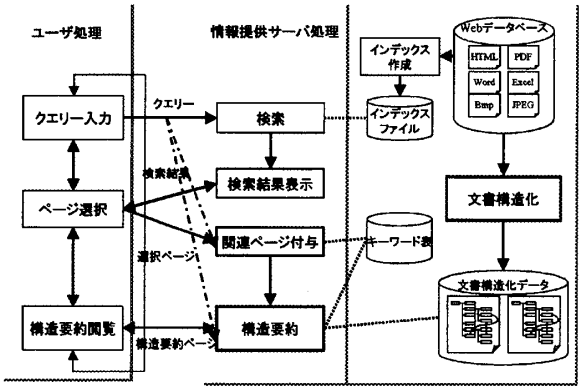

図2　文書構造要約化による情報提供システム

3.2. 文書構造化

文書構造化の処理フローを図3に示す。本処理では，入力された文書データを HTML タグ解析部分，PDF 解析部分，画像解析部分に分割し，各解析処理に振り分ける。次に，各解析処理から得たデータを統合する。そして統合されたデータに言語処理技術を用いて論理構造解析を行い，文書構造化データを生成する。

HTML タグ解析ではタグ情報から構造化を行う。PDF 解析と画像解析では構成要素の抽出手法は異なるが，どちらも2次元的レイアウト情報から構造化を行う。

HTML にはタグ解析のみでは解析不可能な部分や図が含まれる。そのため，HTML タグ解析不可部分は PDF 解析にて処理し，図は画像解析にて処理する。Word，Excel などの文書は PDF に変換し，PDF 解析にて処理する。ただし，PDF 内の図は画像解析にて処理する。JPEG や BMP などの文書画像は，画像解析にて処理する。

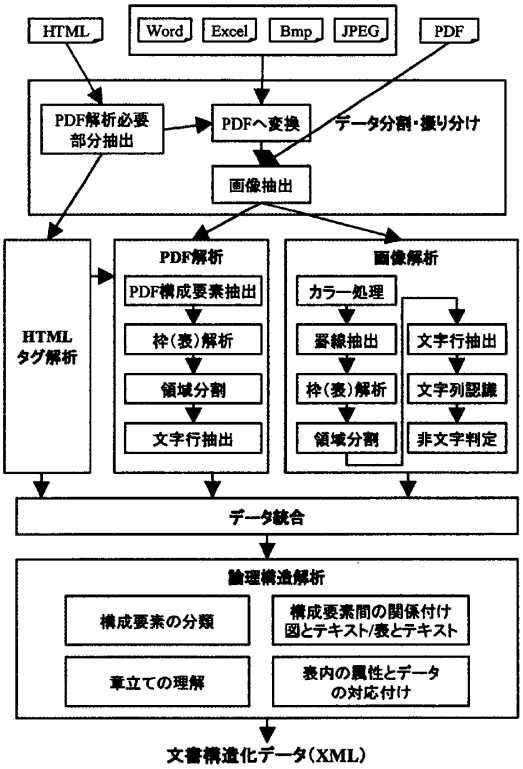

図3　Web データの文書構造化

論理構造解析処理では，構成要素の分類（本文，タイトル，広告など），文書構成要素間の関係付け，章立ての理解，表内の属性とデータの対応付けを行う。これには，形態素解析，構文解析，固有表現技術などの言語処理技術を用いる。

4. 構造要約

構造要約処理では，文書構造化データと検索キーワードを入力とし，ユーザにとって重要な構成要素をそのままに，その他の構成要素を縮小し，2次元的に再配置する。構成要素には，テキスト，図，表，リンク情報がある。これらの構成要素の中から，位置やサイズ、テキスト情報，分類結果，検索クエリー情報を用いて重要な構成要素を抽出する。重要な図については，図中の文字認識結果も利用し，重要なリンクについては，リンク先のリンク件数や，リンク情報と関係する構成要素情報も利用して抽出する。さらに，図・表については重要テキストと関連するものも抽出する。

5. プロトタイプ

検索キーワードに着目した構造要約プロトタイプを作成した。本プロトタイプでは1段組の PDF 文書を対象とし，PDF 解析のみの文書構造化を行う。Adobe(R) Acrobat(R)上で検索から構造要約表示までの処理が実行できる。クリックにより要約部分を縮小／拡大表示切換，要約部分に含まれる図のサムネイル表示，自動目次作成，タイトル部へのジャンプ機能がある。

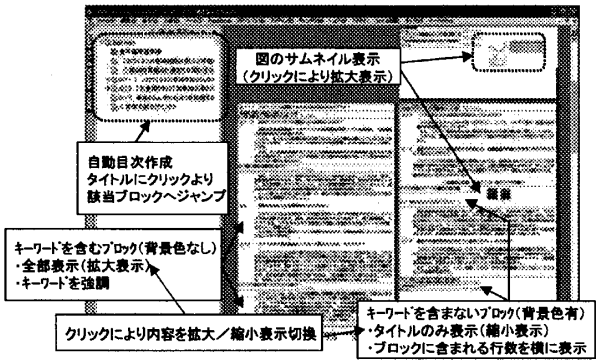

図4　プロトタイプの表示例

6. 結言

本報告では，文書構造要約化を用いて，ユーザが目的の内容に効率的に到達できる情報提供システムを提案し，PDF 文書を対象としたプロトタイプを作成した。

[参考文献]

[1] 品川，“ユーザプロファイルに基づくビューページの動的生成による WWW 閲覧支援”，情報処理学会論文誌,Vol.41 No.SIG6,2000.

[2] 南野，“繰り返し構造に基づいた Web ページの構造化”，情報処理学会論文誌, Vol.45, No.9, 2004.

[3] 石谷，“紙文書を対象としたピボット XML 文書に基づく XML 文書変換システム”,PRMU2003-216,pp.7-12,2004.2.