データベースシステム　122－26
(2000. 7. 27)

モバイル環境における検索エンジンの出力結果の再構成と呈示

近藤宏行† 灘本明代†† 田中克己††

近年，デスクトップコンピュータ上ばかりでなく携帯端末などのモバイル環境において，Web上に存在する既存のコンテンツを提供するためのサービスに注目が置かれるようになってきている．そのような環境では最小限のインタラクションしか行えず，また表示領域が大幅に制限される．モバイル環境で検索エンジンを使って情報検索を行う場合，既存の方法としてはその端末に適した検索エンジンを作成するという方法が取られている．しかし，この方法では各端末に合ったシステムを独自に構築しなければならず，多大な労力要する．さらにカテゴリ型で事前に登録を行うシステムがほとんどである．一般に検索エンジンに問い合わせて得られる結果はキーワードを絞り込まない限り膨大な量になってしまう．これら大量の情報を表示領域が狭く且つ，インタラクションの機会が限られている携帯端末の上で呈示を行うことを考える場合，能動的な呈示よりも受動的視聴型の呈示方式の方が良いと考えられる．そこで，本稿では検索結果の受動的視聴のために適したWebCarrouselを提案する．この方式で，従来の検索エンジンのようにページ単位でランキングを行うのではなく携帯端末で表示するための番組化コンポーネントという概念を用いる．その番組化コンポーネントをカルーセル型で呈示し，ユーザの簡単なインタラクションに応じて次に表示するカルーセルを動的に再構成し呈示する．

Restructuring and Presentation of Results by a Search Engine in Mobile Environment

HIROYUKI KONDO† , AKIYO NADAMOTO†† and KATSUMI TANAKA ††

Recently, much attention has been focused on the realization of several Web information services in mobile phone environments. In the mobile phone environments, both of the opportunities of user interactions and the display area are restricted. This leads to the difficulty of using conventional search engines, and so, many application services are depending on specific built-in search systems. Also, a vast volume of research results are usually obtained by conventional search engines. We believe that the passive viewing of serach results will be appropritae bacause of the narrowness of the display area and the data size of search results. In this paper, we propose a system called WebCarrousel in which research results are organized into several carrousels and each carrousel of Web pages are displayed continuously. By using speech synthesis technology, the search results are presented like as a TV program, which facilitates a passive viewing of search results. Furthermore, for each Web page, the system automatically computes sets of more-detailed, more-abstracted, and similar pages, respectively, and reorganize them as carrousels. Users can navigate among carrousels by simple interaction during passive viewing.

1. はじめに

近年，iモードを始めとする携帯端末において，多様な情報サービスが提供されるようになってきている．その中でも，現在Web上に存在する既存のコンテンツを携帯端末上で提供するためのサービスに注目が置かれるようになってきた．そういった場合，携帯端末では通常のデスクトップPCとは違い，画面の大きさや色数，ネットワークの帯域幅にかなりの制約を負っているので従来のWeb文書をそのまま利用することは難しい．現在のところ何らかの変換サーバを介して携帯端末用にデータの変換を行っているか，又は携帯端末用に作成されたコンテンツを利用するといった場合がほとんどである．だが，今後インターネットに接続可能な端末は更に多様化し，そのためだけのデータ型を作成するよりも，現在ある既存のコンテンツをどのように再利用し有効活用していくかが重要となってくる．更に来年度にはサービスが予定されている次世代携帯端末（IMT2000）では現段階よりはるかにネットワーク帯域幅が向上し，したがって音声読み上げ機能等の高性能な機能が実装することが可能となってきている．本稿ではそのことに着目し，この機能を使った検索エンジンへの問い合わせと，得られた結果の新しい呈示方式を提案する．

一般に検索エンジンに問い合わせ行った場合に得られる結果はキーワードを絞り込まない限りかなりの量になってしまう．これらの大量な情報の中から表示領域が小さく且つ，インタラクションの機会が限られている携帯端末上で呈示を行うことを考えると，能動的な呈示よりも受動的視聴型の呈示方式の方が適していると考えられる．そこで，本稿では検索結果の受動的視聴のためにカルーセル方式での呈示を提案する．受動的視聴を行う場合，その結果がいつ見られたかを特定するよりもあるグループごとにデータを流し続けるカルーセル方式を取ることにより，ユーザが求める情報を発見しや

† 神戸大学大学院 自然科学研究科 情報知能工学専攻
Division of Computer and Systems Engineering, Graduate School of Science and Technology, Kobe University
†† 神戸大学大学院 自然科学研究科 情報メディア科学専攻
Division of Infomation and Media Sciences, Graduate School of Science and Technology, Kobe University

すくする．これを我々は WebCarrousel と呼ぶ．

又，実際に受動的視聴を行う場合インタラクションを最小限に抑えて且つ，ユーザの意図を反映する必要がある．そこで本稿では，あるページを検索結果として得た場合，そのページに対して類似したページを見たいのか，または，そのページよりも詳細なページを見たいのか，または簡潔なページを見たいのか，まったく違ったページを見たいのかの 4 種類のグループに分けることによって次に呈示するカルーセルのグループを動的に再構成する手法を提案する．図 1 にカルーセル型呈示方式のイメージを示す．現在の検索システムではページ単位でランキングを行っているが，カルーセル型提示方式では携帯端末で表示するため単位として番組化コンポーネントという概念を導入する．その番組化コンポーネントをカルーセル型で呈示し，ユーザの簡単なインタラクション（例えば十字キーでの移動）に応じて次に表示する検索結果を動的に再構成する．

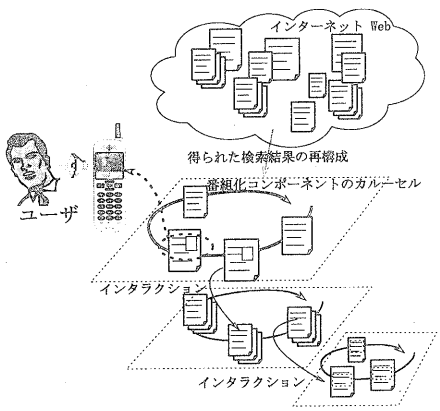

図 1　カルーセル型呈示方式のイメージ

以上より本稿では，

- 番組化コンポーネントを用いた WebCarrousel 生成と呈示
- ユーザインタラクションによる WebCarrousel の動的再構成

という 2 つの事を提案する．以後 2 章では基本的事項と関連研究について，3 章でどのように番組化コンポーネントを導入していくかについて述べ，4 章で検索エンジンに goo を用いたプロトタイプシステムの作成について述べる．5 章で本システムの評価と今後の課題について述べる．

2. 基本的事項と関連研究

2.1 携帯端末での Web 閲覧

現在モバイル端末（携帯端末）上で Web 文書を閲覧するための手段を下記に記す．

- Compact HTML
  - NTT Docomo「i モード」
- WAP（Wireless Application Protocol）
  - DDI セルラー「EZ サービス」
- MML（Mobile Markup Language）
  - MOBiDY プロジェクト 慶応大学 SFC 研究所 等

Compact HTML は従来の HTML の機能を使った，携帯電話や PDA などハンディタイプ端末用マークアップ言語で従来の HTML のタグやオプションを制限したものである．Compact HTML の利点は，独自のサーバを用意する必要がなく既存の Web サーバを利用し，既に膨大な数ある既存の Web 文書を利用することができることである．しかし実際に検索を行う場合，i モードに対応するよう，ほとんどが yahoo のようなカテゴリ型の登録制サーチエンジンで，サイト自体も Compact HTML で記述されている．

それに対し，WAP とはマークアップ言語からプロトコルまで全てを携帯端末用に最適化するように考えられた方式である．WAP はゲートウェイと呼ばれる専用の装置を WAP クライアント（携帯端末）向けのインターネットへのアクセスポイントに設置し，携帯端末の特性に最適化された WAP プロトコルにおいて通信が行われる．有線区間では従来のインターネットの標準プロトコルに基づいた通信が行われ，ゲートウェイはその間でプロトコルおよびコンテンツフォーマットの変換などを行う．そのため，コンテンツの記述には HDML(Handheld Device Markup Language) を用い，また次世代の WML(XML に基づいた携帯電話用ファイルのタグ言語) 言語も検討されている．WML はカードとデッキという概念を持ち，カードとはブラウザ 1 画面に表示されるコンテンツであり，カードの集合がデッキである．一つの WML ファイルはデッキであり 1 度にダウンロードされる．

MML は，Web 上で利用されている HTML を，移動体通信網へ適応させるために簡略化した，Application Layer の Language である．MML では基本的に HTML で使われているタグが短縮して使用されるため，MML は HTML のサブセットに位置づけられる．

次世代携帯端末の通信方式では，現在よりも大幅に帯域幅が向上し通信速度が増すに伴い，音声読み上げ機能などの実装も十分に行える環境になってきている．

また既存の Web コンテンツを利用するため方法として慶応大学萩尾研究室 [1] では，携帯電話に限らず，その他のモバイル端末例えばカーナビや PDA において統一したコンテンツの提供を行うために，独自のプロキシの実装が行われている．このプロキシは一般の Web コンテンツを各端末のブラウザの機器能力に応じて変換を行い，これにより表示画面の大きさや色数の制限に柔軟に対処できるように作成されている．また，HTML を携帯コンテンツに変換するサービスとして，FLEX FIRM 社 [2] の X-Servlet(クロスサーブレット)，Spyglass 社 [3] から Prism という製品等が発表されている．Prism は携帯が対象デバイスの場合，コンテンツをそのまま表示させることが困難であるので Data Extruction（任意のデータ抽出）を用いた変換を行っている．

2.2 Web の受動的視聴

我々はこれまで，膨大な Web 情報を容易に取得することを目的とし Watch and Listen 型インタフェースとして，受動的視聴型番組呈示方式を提唱してきた．[4)~7)] これらは，テキストや画像で表示されていたコンテンツを，音声やキャラクターアニメーションを用いて表現することで，これまでの

Webブラウザと異なり，インタラクションを極力抑えた上で有効に情報を提供できる環境を提供している．そこでは，レイアウト情報として与えられている，テキストと他のコンテンツとの時間軸的同期情報をどのように扱うかということに重点をおいている．これらの同期情報を扱うための番組化の技術を，携帯端末においても利用することは有効であると考えられる．

2.3 WebSkimming

角谷ら[8]は得られた検索結果から，そのページの共通の親ページを発見することにより検索結果に一連の流れを持たせストリーム化して呈示する方法を提案している．WebSkimmingでは幾つかのページを内容によってシリアライズすることで，単一のページを見るより分かりやすくユーザに呈示しようと試みられている．ページごとの関連度により次に見せるページの演出（例えばワイプアウトしたり，スライドインする）を変える呈示方法を取っているが，実際にはどの順番で見せるかということに重点が置かれていてあまりインタラクションをするという部分には触れられていない．

2.4 意味的な単位を考慮したWeb検索

近年，検索エンジンに関する研究は様々な角度から行われるようになり，例えば，Webページを単体として扱うのではなく，意味的なまとまりからページ群を基本単位として扱うといったことも行われている．田島ら[9]は，Webページの検索単位として，Webページ単体ではなく，Webページのリンク構造から静的に抽出した「カット」と呼ばれる意味的なまとまりを検索の単位とすることを提案している．また，波多野ら[10]はWeb上のハイパーテキストデータを意味的につながった論理的な文書の単位として「極小マッチ部分グラフ」を提唱し，それらを基本単位とした検索機構の提案および実装を行っている．これらの研究では，ユーザが検索機構に問い合わせを行った際に検索結果が動的に得られるという特徴をもっている．

また，更に文書自体の意味的な情報ばかりでなく構造的な情報，特にリンク情報を用いた検索エンジンの実装も多数なされている．実際にはclever[11]やgoogle[12]といったポータルサイトにおいて，hub（信頼できるリンク集）とauthority（問い合わせに対し的確だと推測されるページ）という2つの概念をページに与えることにより，検索の精度を上げようと従来の意味的な類似度ばかりからではなく，構造情報を元にした検索手法の改良が行われている．

本稿ではページの集合ではなくページの部分から成る集合という観点から検索結果を分類し呈示する手法を考える．

3. 携帯端末における検索結果の呈示方式

携帯端末において実際のWebページを閲覧しようとした場合，従来のコンピュータ環境と比べ以下の二つの制限があると考えられる．

- コンピュータ上で見る場合に比べて一度に表示可能な情報が圧倒的に制限される
- インタラクションの機会もかなり制限される

実際にコンピュータ上でWebの検索を行う場合，得られるのは単なる結果のランキングであるが，これをそのまま携帯端末上で上から順番に表示していくだけでは，膨大な検索結果が得られた場合，上記の制限のためなかなか求める情報に辿り着く事は難しい．

そこで，我々は実際に検索エンジンにより得られた検索結果の呈示方式として，簡単なインタラクションを伴ったカルーセル型呈示方式を提案する．図1はそのイメージを示している．携帯端末で検索結果を表示するための最小構成要素として番組化コンポーネントを定義する．これらの番組化コンポーネントの組み合わせを携帯端末に次々と検索結果をカルーセル型で送っていくことを考える．更に，現時点で視聴しているページに対し，それに類似したものが見たい，もっと詳細なものが見たい，もっと簡潔なものが見たい，違うものが見たいという4つの簡単なインタラクションを行うことで，それぞれのグループ分けから新しいカルーセルを生成し，それをまた順番に視聴するWebCarrouselを提案する．

以上のことから，本稿では検索結果の中から求めるページを閲覧するために下記の4つのフェーズを行うことを提案する．

(1) 番組化コンポーネントの作成
(2) 番組化コンポーネントの組み合わせのカルーセルグループの作成
(3) 番組化コンポーネントのカルーセル方式での呈示
(4) ユーザのインタラクションによる次に呈示するコンポーネントおよびカルーセルグループの動的再構成

3.1 番組化コンポーネント

我々は，携帯端末上で表示するための単位として番組化コンポーネントCというものを提案する．番組化コンポーネントとはカルーセルを構成する一つ一つの単位である．また番組化コンポーネントとは，ある検索結果ページPを最も端的に表すためのPの1部分であると考えられる．番組化コンポーネントは

$$C = (c_{head}, c_{main}, c_{sync})$$

のように3つの部分から構成される．それぞれc_{head}はPからタイトル，見出しを抜き出し作成したヘッダコンポーネント，c_{main}はページ中の一部分を抜き出したメインコンポーネントであり，実際には音声にて呈示される部分である．c_{sync}はそのメインコンポーネントに対する同期化可能領域とする．同期化可能領域とはあるHTML文書が与えられた場合，その文書の構造から推定した同期している領域のことである．[7] c_{symc}は音声で呈示されているc_{main}と同時に端末上に表示するものであり，本稿では画像とアンカーのみを考慮する．更にそれぞれの要素について空，又は複数の可能性もある．

以下，HTML文書のタグはその性質上4つの種類に分けられると考える．

- 構造タグ S_{tag} ($H1 \sim H6$, P, $BLOCKQUOTE$, DIV, VL, OL, DL, $TABLE$)
  これらのタグは主に文書構造を示すために用いられる．一般にHTMLから文書構造を抜き出すためにこれらのタグを利用する．
- 強調タグ ($STRONG$, EM, TT, I, U, B, BIG, $SMALL$, $STRIKE$, S, $FONT$)
  これらのタグは主に強調などの特殊な表示を行ったりす

る時に用いられる．これらは文書論理構造を直接示すわけではないので領域を推定する場合は無視できる．

- 埋め込みタグ E_{tag} ($IMAGE$, A, $FORM$, $APPLET$, $OBJECT$, $EMBED$, MAP)
  これらのタグは画像や，アンカーといった他の特殊なオブジェクトを埋め込むために使用する．これらは主に同期化可能なオブジェクトとして番組化コンポーネントに組み込むことを考える．またこれらのタグの中に構造タグが入るということは無いので，この場合必ず終端にくる．
- その他のタグ
  その他のタグは番組化コンポーネントを作成場合重要ではないので無視する．

番組化コンポーネントを作成するためには，ページを構造から幾つかの部分に分割した最小構成単位 $p_i, i \in (1,..,m)$ を考えておく必要がある．HTML 文書から論理木を導き出すことにより構造の解析を行う手法は色々と提案されている．[13),14)] また我々はこれまで Web 情報の番組化として，HTML 文書の構造情報から同期化可能領域を発見する手法を幾つか提案してきた．[6),7)]

以上のことを踏まえてて番組化コンポーネントの作成手順を以下に示す．また，図 2 では，HTML 文書をエッジラベル付きグラフと考えて作成手順の概要を示している．

ヘッダコンポーネント c_{head} の抽出

まず，ヘッダの部分と本文の部分に分割し，ヘッダの部分からタイトルを抜き出し一つの p_{head} とする．

メインコンポーネント c_{main} の候補探索

次に，c_{main} を本文中から抜き出す必要がある．一般的に Web 検索を行う場合幾つかのキーワードを利用するが，検索キーワードを $k_i, i \in (1,..,n)$ とする．まず，その文書の中で検索キーワード k_i を含む，文章を探す．文書の頭から順番に検索していき，最初に見つけた文章を含む極小な範囲のタグ領域 $p_k, k \in i$ を見つけ，それを c_{main} の候補とする．その場合，直接的に文書の構造を決定する構造タグのみを利用し，それ以外は無視する．

メインコンポーネント候補 p_k の修正

候補となった部分の文章が大きすぎる場合や，又は逆に小さすぎる場合その範囲修正を行う必要がある．

- まず，大きすぎた場合はその下に構造タグを含む場合，その構造タグを切り離す．そのタグが末端である場合はキーワード k_i を含む文章のみを抽出する．
- 小さすぎる場合は，その極小な範囲のタグの親のタグまで戻ってそれを一つのコンポーネント候補 p_k として定義する．
- H タグで囲まれている場合，その部分は c_{head} の候補になると考える．故に，その場合は一つしたのタグを含める．

以上をその文書中に k_i が出現している間繰り返し，検索キーワードを含む幾つかの p_k がページから抜き出され作成される．

候補からの選択

次にそれら p_k の中からメインコンポーネントを選択する必要があるが，次の手順で選択，抽出を行うことを考える．

その文書内における重みつき $\mathbf{tf} \cdot \mathbf{idf}$ でそのページにおいて p_k がどの程度検索キーワード k_i を表現しているかを計算する．重み付きとは，キーワードが強調タグで囲まれている場合，それが含まれている文章の方がよりメインコンポーネント候補として適していると考えられる．このときある p_k におけるキーワード k_i の重み w_i は以下のように定義し，$\sum w_i$ の一番大きな p_k から順番にそのページ内におけるランクをつける．(α は強調タグによる重み，N_i はその文書中の単語数)

$$w_i = \mathbf{tf}(\alpha \cdot k_i) \cdot log\frac{N_i}{\mathbf{df}(k_i)}$$

これより，p_k をランク順にいくつか採用し番組化コンポーネントのメインコンポーネントとする．

番組化コンポーネントの生成

以上より，メインコンポーネント c_{main} を決定する．c_{main} に c_{head} 候補の見出しが入っていればそれを，c_{head} に付加する．次に c_{main} として決定された領域 p_k に対応する画像等の同期化可能領域を付加し，この領域を同期化コンポーネント c_{sync} とする．同期化可能領域を求めるために我々が従来考案した手法 [7)] を用いる．このようにして決定した c_{head}，c_{main}，c_{sync} をまとめて，同期化コンポーネント C とする．

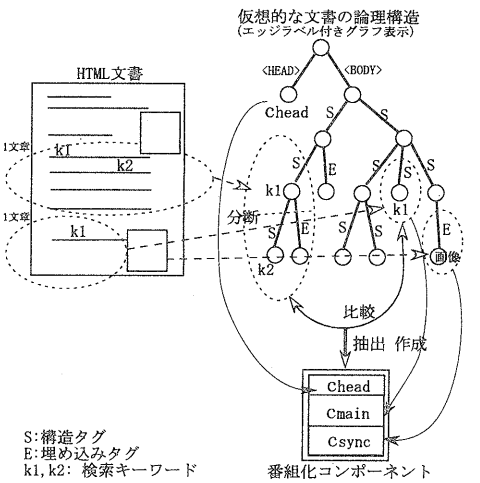

図 2　番組化コンポーネントの作成

以上のことから番組化コンポーネント C はあるページの一部分を抽出した形であると言える．最小構成要素の候補となった部分を $c_i, i \in (1,..,n)$ とすると，図 3 のように表される．その場合構成要素候補のランキングが，このページ内のトップレベルの情報からより詳細な情報までを表しているとも言える．例えば，見出しの部分は上位のレベルの情報，更に本文にはより詳細な情報であるとも仮定できる．

3.2　WebCarrousel の作成と呈示方法

WebCarrousel とは，問い合わせの結果得られた文書をグループ分けしたものだと考えられる．まず最初のカルーセル

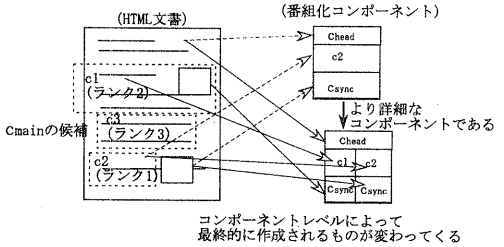

図 3　番組化コンポーネントのレベル

として，検索エンジンに問い合わせを行いランキング情報を元に上位の何件かを取り出し，最初のカルーセルとして呈示する．

そこで，実際に呈示を行う場合において，これまで述べてきた番組化コンポーネントの構成レベルによって，それぞれの呈示方法が異なってくる．

3.2.1　基本的な呈示方法

まず，番組化コンポーネントが c_{head}，c_{main}，c_{sync} 全てが一つずつ組になっている場合，最もシンプルな画面レイアウトを用いることが可能であり，以下に呈示の流れを行示す．

(1)　c_{head} があればそれを表示．
(2)　c_{main} のパラグラフの音声読み上げを行う．
(3)　同時に c_{sync} が存在すればそれを画面で表示する．
(4)　その時点で，ユーザからのインタラクションがなければ，そのページの次の番組化コンポーネントの呈示へ移る．

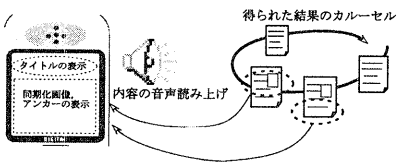

図 4　番組化コンポーネントの呈示イメージ

基本的には以上であるが，番組コンポーネントの構成上もう少し考慮しなければならない事項がある．まず，各要素が空の場合 (c_{main} 以外) 表示は行わない．次に，c_{main} が二つ以上から構成される場合，コンポーネント作成の時点で c_{main} の構成候補である p_k にはランク付けを行っているので，その順番にパラグラフの再生，同期領域の呈示を行う．更に，現在再生中の c_{main} に対して，複数の c_{sync} が存在する場合それらを同時に表示するのか，又は順番に表示するのかが問題となる．本稿では同期化オブジェクトとして，イメージ，アンカーのみを考慮しているので，同種類のオブジェクトの場合はそれぞれ元々のレイアウト順にシリアライズし，異なる場合は同時に表示することを考える．

3.2.2　ページ全体の呈示方法

ユーザが番組化コンポーネントのカルーセルを閲覧し，最終的に見たいページに行き当たった場合，そのページ自体をどのように呈示するかということも問題となってくる．現在，i モードで HTML 文書を閲覧しようとした場合，ある一定のサイズ以上はまったく表示されない．よって，そのまま HTML 文書を呈示するよりも，以前に定義したように，番組化コンポーネントはページ自体のサブセットであるので，番組化コンポーネントの呈示方式を拡張する形を取る．

まず，HTML 文書中において，検索キーワードを含まない部分は c_{main} の候補として挙げられない．その部分について，同様に極小なタグ領域を求め $p_l, l \in n$ とする．そのように考えるとページ全体は p_k と p_l の集合から構成される．HTML 文書は順序つきの構造になっているので，構造的にトップレベルの p_k 又は p_l から同様の方式で呈示を行っていく．

3.3　ユーザインタラクションによる WebCarrousel の再構成

ユーザはそのページの番組化コンポーネント，又はページそのものを閲覧した場合において簡単なインタラクションを行い，そのインタラクションの種類に応じて次に閲覧するカルーセルを決定する．そのためにユーザが実行可能なインタラクションとして以下の 4 つの評価基準を考える．図 5 にインタラクションのイメージを示す．

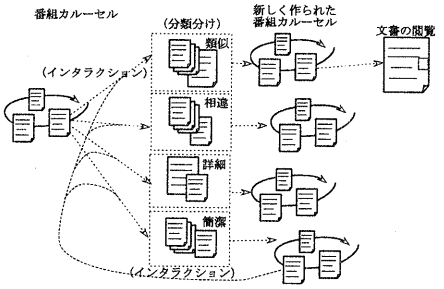

図 5　ユーザインタラクションのイメージ

- **類似**：現在見ている C に類似した情報を要求する場合
- **相違**：現在とは違った話題の情報を要求する場合
- **詳細**：現在見ている番組化コンポーネント C よりももっと詳細な情報を要求する場合
- **簡潔**：現在見ている C よりも，もっと簡潔な情報を要求する場合

これらを区分けする場合，

- 文書同士の特徴量を基準にした分類
- 文書内，又は文書間の構造を基準にした分類

という二つの違った観点からグループ分けを行うことが可能であると考えられる．特徴量を元にするとは，文書間の意味的な繋がりからグループ分けを行うことを考えたものである．

3.3.1　特徴量の類似性によるタイプの判別

実際にインタラクションを行う場合において，類似，相違，詳細，簡潔の評価基準をどのように決めるかが問題になってくる．本節では，類似，相違，詳細，簡潔といったことを特徴ベクトルの形から推測する手法を提案し，実験を行う．これらを判別するためには様々な方法が考えられるが，まず一番シンプルな方法から試行する．

最初に，Web 文書 P_i における特徴ベクトルを出現単語 $w_1, \cdots, w_n$ を各基底とする重要度 $W_1, \cdots, W_n$ として求める

と，各文書の特徴ベクトル $\mathbb{F}(P_i)$ は

$$\mathbb{F}(P_i) = (W_1^i, \cdots, W_n^i) = \frac{1}{N_i}(f_1^i, \cdots, f_n^i)$$

で表される．ここで，$f_j^i, j \in (1, \cdots, n)$ は Web 文書 P_i における単語 w_j の出現回数，N_i は P_i の総単語数を表す．即ち W_j は単語の出現頻度 $\mathbf{tf}$ を表している．

2 つの文書間についての特徴量を考えるが，以下のようなアルゴリズムを提案する．ここでは，文書 P_0 を基準として文書 P_1 における類似度，相違度，詳細度，簡潔度をそれぞれ計算し，P_1 がどのタイプにもっとも当てはまるかを求め分類を行う．図 6 にこれらの典型的な特徴ベクトルの形を示す．

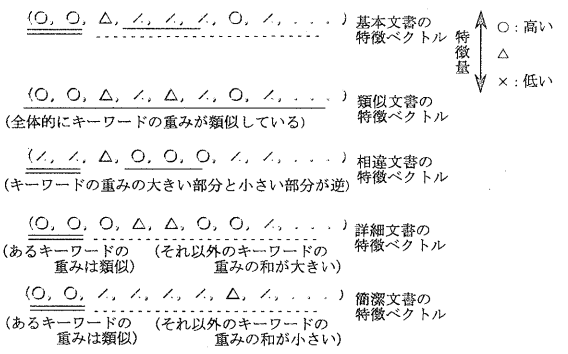

図 6 特徴ベクトルから仮定する類似，相違，詳細，簡潔

(1) まず，2 つの文書間の類似度を λ_{sim} とすると，λ_{sim} は比較する 2 つの文書 P_0，P_1 の特徴ベクトルのコサイン相関値 $S(P_0, P_1)$ より求める．

$$\lambda_{sim} = \mathbb{S}(P_0, P_1) = \frac{\mathbb{F}(P_0) \cdot \mathbb{F}(P_1)}{||\mathbb{F}(P_0)|| \cdot ||\mathbb{F}(P_1)||}$$

λ_{sim} の値が高ければこれは類似文書と考えられる可能性が高い．

(2) 相違度を考慮する場合，2 つの文書が相違しているとは文書の特徴量を比較して，基本文書の特徴量が高い部分については比較文書の特徴量が低く，逆に基本文書の特徴量が低い部分については比較文書の特徴量が高くなっていると仮定できる．よって，2 つの文書間の非類似度と考えられる．相違度を λ_{diff} とすると，上述のコサイン相関値より，

$$\lambda_{diff} = 1 - \mathbb{S}(P_0, P_1)$$

と考えられる．λ_{diff} の値が高ければ，話題の違う文書である可能性が高い．

(3) 詳細度を求める場合，基準の文書よりも詳細な文書というものは基準よりも多くの付加情報をもっていると仮定できる．その付加情報は特徴ベクトルで表した場合，ベクトルの類似している部分以外の差がどれだけあるかいうことで表されると考えられる．そこで，2 つの特徴ベクトルの要素を各一つずつ比べた場合，1 つ以上の要素が類似している，すなわち $||W_j^0 - W_j^1|| < \theta$ (θ は閾値) である時，その要素以外の部分で $\mathbf{tf}$ を計算した新しい特徴ベクトル $\mathbb{F}'(P_1) = (W_1^{1'}, \cdots, W_n^{1'})$ を作成する．この場合 P_1 の詳細度を求めるので，同様の手順で P_0' を計算する必要は無い．図 7 に概念を示す．$\mathbb{F}'(P_1)$ から $\mathbb{F}(P_0)$ を引いた差を 特徴ベクトルの次元 n で平均を取ったものを詳細度 $\lambda_{detail} = \mathbb{D}(P_0, P_1)$ とすると，

$$\mathbb{D}(P_0, P_1) = \begin{cases} \frac{1}{n}\sum_{j=1}^{n} (W_j^{1'} - W_j^0) & W_j^{1'} > W_j^0 \\ 0 & other \end{cases}$$

以上より，λ_{detail} が高い文書は注目すべきキーワードの他に付加情報を多く持つと考えられ，より詳細な情報として位置付ける．

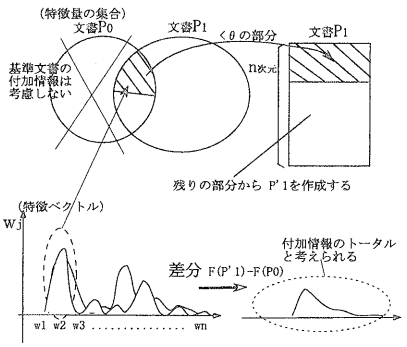

図 7 特徴量を元にした 2 文書間の詳細関係

(4) 簡潔度を求める場合，簡潔であるというのは詳細の逆，すなわちあるキーワードの情報に対しては同程度の特徴量を有するが，その他の部分の特徴量が少ない場合であると考える．即ち上記と同様に $||W_j^1 - W_j^2|| < \theta$ である時，今度は逆に $\mathbb{F}'(P_0)$ を生成し，$\mathbb{F}'(P_0)$ から $\mathbb{F}(P_1)$ を引いた差を 特徴ベクトルの次元 n で平均を取ったものを簡潔度 $\lambda_{brief} = \mathbb{B}(P_0, P_1)$ とすると，

$$\mathbb{B}(P_0, P_1) = \begin{cases} \frac{1}{n}\sum_{j=1}^{n} (W_j^{0'} - W_j^1) & W_j^{0'} > W_j^1 \\ 0 & other \end{cases}$$

以上より，λ_{brief} が高い文書は注目すべきキーワードの他に付加する情報が基準文書に比べ少ないと考えられ，より簡潔な情報として位置付ける．

以上の手順で類似度，相違度，詳細度，簡潔度を求めそれらが高いもの数件を抜き出すことで，次に呈示するカルーセルを決定する．

上記の手法の有用性を確かめるために，実験を行った．まず，初めに検索エンジンに問い合わせを行い，得られた結果の中から類似，詳細，簡潔といった関係にある文書を幾つか事前に用意する．相違関係については対象文書数を増やすにつれて非常に高い値を示すものが多く現れてくるので，この評価は行なっていない．次に得られる検索結果を順番に増やしていく時，カルーセルを構成するために選出されるのは類似度，詳細度，簡潔度が高い順であるので，それら値の高い上

位何ページかを抜き出しシステムの解とする．表1では，そのシステムの解の内に事前に類似関係，詳細関係といったものを調べて，用意した正解がいくつ含まれているかの割合を示している．今回は，検索結果に100件から500件，基準となる文書を5つ選び各文書における正解の出現率の平均値を出している．ここでは，ランキング上位10件をシステムの解として採用した．

又，図9に検索ページ500ページとした場合に，システムの解として10件，20件，50件，100件を採用した時の適合率と，再現率を示す．

類似度については従来の **tf** を用いたコサイン相関値を用いたのでそれなりの値を示した．又，今回提案した詳細度についても約50%の割合で正解得ることが可能である．しかし逆の簡潔度についてはあまりあてはまっていない．また，詳細な文書であっても文書量が大きくなりすぎた場合，相違する文書の方へ偏ってしまうという結果も得られた．

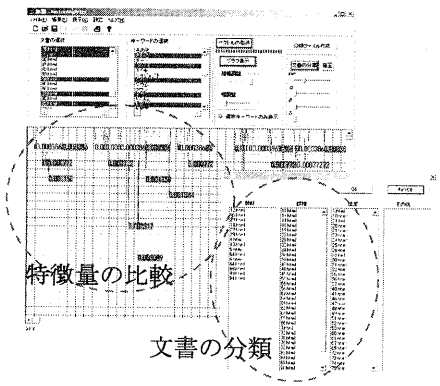

図 8 特徴量を用いた分類判定実験

検索結果数	類似文書	相違文書	詳細文書	簡潔文書
100 件	90	-	73	30
200 件	82	-	70	30
300 件	82	-	59	21
500 件	73	-	45	17

表 1 実験結果（各分類の正解の出現率（適合率）：単位は%）

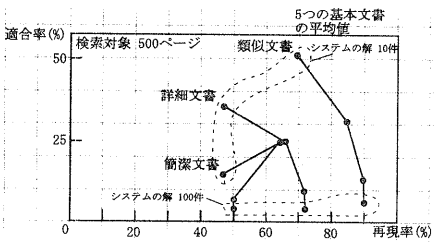

図 9 類似度，詳細度，簡潔度の適合率と再現率

本実験で用いた特徴量から類似，相違，詳細，簡潔に分類する手法は一番手始めのものであり，まだまだ様々なバリエーションが考えられる．平田ら[15]はその文書に出現する単語量と、単語の共起関係を利用してマッピングを行い，これらの分類を行う提案，実験を行っている．これらの分類をより効果的に行うためには様々な点を改良し実験を重ねていく必要がある．

3.3.2 文書内及び，文書間の構造によるタイプの判別

前述の特徴量を用いた分類法は，文書の意味的な内容から類似文書，詳細文書等を推測したものである．それとは逆に文書の構造，又はリンク情報からそれらの分類を行うことも考えられる．例えば，1つの文書におけるタイトルや見出しは言わばトップレベルの情報であり，下に行くにつれてより詳しくその内容について書かれていると考えられる．また，リンク集などを考えた場合，リンク元がトップレベルの情報となりリンク先に詳細なことが書かれてある可能性が大きい．このように構造情報から分類する方法は特徴量を用いるものとは別の次元で考慮する必要がある．しかし，現在検討中なので本稿では触れていない．

4. プロトタイプシステム

前章の案を基にして，デスクトップコンピュータ上で実際にプロトタイプシステムの実装を行う．本システムはすべてWindowsプラットフォーム上で作成する．主な開発環境としてMicrosoft VisualC++を使用した．実際の実装イメージを図10に示す．

本プロトタイプシステムではまずユーザからの問い合わせがあった時，メインコントローラを介して検索エンジンに問い合わせを行う．今回は検索エンジンにgoo[16]を用いた．得られた結果を元にそれらを番組化コンポーネントに分割し，さらにそのコンポーネントを分類し番組カルーセルを作成する．これらの処理はすべてメインコントローラ上で行う．そのあと，送り出す順番に携帯端末で見られる形式に変換し，クライアントである端末へ送信する．

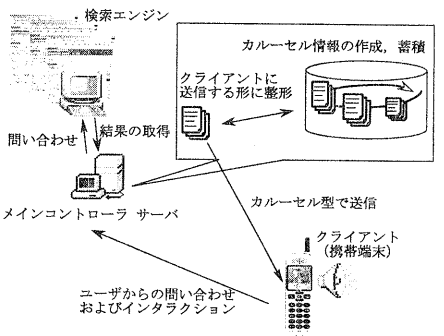

図 10 プロトタイプシステム概要

今回はプロトタイプシステムとして，実際の携帯端末ではなくデスクトップコンピュータ上にシミュレーションとして，簡単な音声読み上げブラウザを作成する．(図11は開発中の画面例) クライアント側では送られてきた，番組化コンポーネントについて，ヘッダコンポーネントと同期化コンポーネントを画面上で表示しながら音声読み上げ機能によりメインコンポー

ネントの部分を呈示する．音声読み上げ機能には Microsoft Speech API および NEC SmartVoice の音声合成エンジンを利用した．又，サイズの大きな画像は縮小してクライアントに送る．そのあと，4 つの簡単なインタラクションをユーザが行うことによって次に表示するコンポーネントが動的に再構成されるようになっている．又，一度問い合わせが行われた検索キーワードについては次回から同じ計算をするよりも，インデキシングすることで処理速度を上げる必要もある．このように，本システムはまだ現在開発中であり，更に適切なインタラクションの改良などを施す必要がある．又，計算機上のシミュレーションではなく実際の携帯環境で使用に耐えうるかどうかの実験を更に行っていく必要がある．

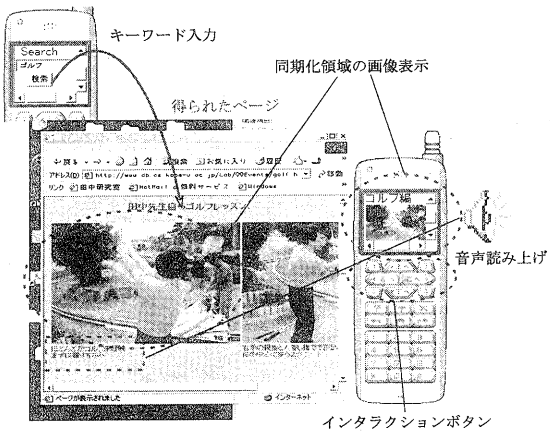

図 11 プロトタイプシステム（クライアント）

5. おわりに

本稿では，携帯端末という様々な制約を受ける環境で，既存の Web 文書を閲覧し，また検索結果の新たな呈示方式，WebCarrousel を用いる手法を提案した．携帯端末環境では一度に表示可能な領域がかなり制限され，またインタラクションの機会も限られてくる．これらの制限がある為，デスクトップコンピュータ上での能動的な呈示方法よりも寧ろ受動的な呈示方法が適している．受動的な呈示方法であるために，ユーザがそのページをきちんと見るとは限らないので，各ページからそのページを最も端的に表現している番組化コンポーネントを抽出し，それらをグループ化してカルーセル型での呈示を用いた．更に，最小限のインタラクションとして，現在見ているページを基準にしその他のページを類似，相違，詳細，簡潔の 4 つのグループに分け，それをもとに新たなカルーセルを構成し，よりユーザが求めるページ探索を支援する．そのために文書間の特徴量を用いた一つの仮説の方法を用い実験を行った．

本稿では，番組化コンポーネントを定義したが，今回は単一のページを検索結果の対象と考えた．しかし，検索結果の集合を更に番組カルーセルの単位として考えることも可能である．また本稿で用いた特徴量から類似，相違，詳細，簡潔に分類する手法は一番手始めのものであり，まだまだ，様々なバリエーションが考えられる．そのために様々な点を改良し実験を重ねていく必要がある．

謝辞 本研究は一部，文部省科学研究費「分散型ハイパーメディアからの構造発見とアクセス管理」(課題番号 12680416)，および日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」(プロジェクト番号 JSPS-RFTF97P00501) によっております．ここに記して謝意を表すものとします．

参考文献

1) 平川泰之, 遠山緑生, 安田絹子, 石川雅康, 浅田卓哉, 服部隆志, 萩野達也: 次世代モバイル Web システム, 産学官交流シンポジウム YRP 移動体通信産学官交流シンポジウム 2000.
2) FLEXFIRM: http://www.x-servlet.com/.
3) Spyglass: http://www.spyglass.co.jp/.
4) 近藤宏行, 角谷和俊, 田中克己: 番組メタファーを用いた情報検索結果の提示方式, 情報処理学会研究報告, 99-DBS-119-70, pp. 419–424 (1999).
5) 服部多栄子, 角谷和俊, 灘本明代, 草原真知子, 田中克己: 番組メタファーによる Web ページの利用者適応型呈示方式, 情報処理学会研究報告, 99-DBS-119-69, pp. 413–418 (1999).
6) 灘本明代, 服部多栄子, 近藤宏行, 沢中郁夫, 草原真知子, 田中克己: Web 情報の番組化のためのオーサリング機構, 情報処理学会研究報告, 00-DBS-120-14, pp. 99–106 (2000).
7) 服部多栄子, 沢中郁夫, 灘本明代, 田中克己: Web の受動的視聴のための同期化可能領域の発見と番組化用マークアップ言語 S-XML, 情報処理学会研究報告, 00-DBS-121-2, pp. 9–16 (2000).
8) 角谷和俊, 正賀信寛, 上原邦明: WebSkimming WWW ページ群の動的要約による閲覧支援, 電子情報通信学会データ工学ワークショップ (DEWS'2000) 論文集 (2000).
9) Tajima, K., Hatano, K., Matsukura, T., Sano, R. and Tanaka., K.: Discovery and Retrieval of Information Units in Web, *Proc. of the Workshop on Organizing Web Space (WOWS'99) in conjunction with ACM Digital Libraries* (1999).
10) 波多野賢治, 佐野綾一, 段一為, 田中克己: 自己組織化マップと検索エンジンを用いた Web 文書の分類ビュー機構, 情報処理学会論文誌: データベース, Vol. 40, No. 1, pp. 933–942 (1999).
11) Kleinberg, J. M.: Authoritative Sources in a Hyperlinked Environment, IBM Research Report RJ 10076(91892) May 1997.
12) Google.Inc.: http://www.google.com/.
13) 品川徳秀, 北川博之: ユーザ視点に即した仮想 WWW ページの動的生成による閲覧支援, 情報処理学会研究報告, 99-DBS-119-71, pp. 425–430 (1999).
14) N.Ashish and C.A.Knoblock: Wrapper Generation for Semi-Structured Internet Sources, *ACM SIGMOD Records*, Vol.26,No.4, pp. 8–15 (1999).
15) 平田陽一, 松倉健志, 田島敬史, 田中克己: Web 検索における意味的適合フィードバック機構, 情報処理学会研究報告, 00-DBS-122-21 (2000).
16) NTT ヒューマンインターフェース研究所: goo パワーサーチ, http://www.goo.ne.jp/.