コンテンツ間距離の標準偏差に基づくWebページ動的分割方式

服部 元[†] 松本一則[†] 菅谷史昭[†]

携帯電話を利用してインターネット上のPC向けのWebページを効率的に閲覧する要求が高まっている．しかしながら，Webページのほとんどは PC で閲覧することを想定して作成されており，携帯電話で閲覧可能な Web ページはごく一部である．そのため Web ページ中の個々の情報を分割して階層的にする等，情報量の多い Web ページを携帯電話向けに再構築する必要がある．我々はこれまで，アンカや写真等のコンテンツの間にある HTML タグの深さと数を利用して算出したコンテンツ間距離に基づく Web ページ自動分割方式について検討し，携帯電話の小画面でも PC 向けの Web ページを効率的に閲覧可能とした．しかしながら，Web ページによりコンテンツ間距離の分布が大きく異なるため，すべての Web ページに最適となる分割閾値を固定的に設定することが困難であった．そこで本稿では，Web ページごとに適切な分割閾値をコンテンツ間距離の標準偏差に基づき動的に決定する方式を提案する．既存の方式と比較する評価実験を行い，有効性を確認した．

Dynamic Segmentation of a Web Page Based on Standard Deviation of Content-distances

Gen Hattori,[†] Kazunori Matsumoto[†] and Fumiaki Sugaya[†]

The demand of retrieving information from general Web pages using a mobile phone is increasing. However, since most of the general Web pages in the Internet are produced for PC users, mobile phone users cannot retrieve plenty information from Web. Therefore, we have to divide a Web page into articles and rebuild the tree structure of Web pages. In the previous work, we examined the automatic Web page segmentation scheme which uses the distance between contents based on the relative HTML tag hierarchy, which is the number and depth of HTML tags in Web pages. The system makes it possible to display a PC-oriented Web page on the small screen of a cellular phone efficiently. However, since the distribution of the distance between contents was unique for every Web page, it was difficult for us to set a suitable threshold for segmentation of all Web pages. In this paper, we propose a new dynamic scheme to dicide the threshold for every Web page, based on the standard deviation of the distance between adjacent contents. We conduct the evaluation experiment in comparison with the existing schemes, and show the effectiveness of our proposed scheme.

1. はじめに

日本の携帯電話は 2005 年 8 月時点で 7600 万台以上が Web ブラウザ機能を搭載しており[18)]，多くの携帯電話ユーザは鉄道の乗り換え案内や天気予報，最新ニュース等の様々な情報を得ることができる環境にある．しかしながら，携帯電話の場合は小さな画面や自由度の低い入力デバイス等，ユーザインタフェースの制限があることから，携帯電話向けの Web ページが持つ情報量は少なく調整される傾向にあり，さらに携帯電話向けの Web ページの数は PC 向けと比較してはるかに少数である．そのため，携帯電話を利用して情報量が豊富な PC 向けの Web ページを閲覧する要求が高まっている．実際に，現在 PC 向けの Web ページを表示可能なフルブラウザを搭載した携帯電話端末が市販されているが，PC 向けの Web ページは一般的な Web ブラウザを対象とした画面のレイアウトや情報量に調整されており，比較的大きな画面とキーボードやマウス等の自由度の高いユーザインタフェースを利用しなければ容易に閲覧することはできない．そのため携帯電話で PC 向けの Web ページを容易に閲覧するためには，Web ページを携帯電話向けに再構築する必要がある．

Web ページを携帯向けに再構築する方式として，我々はこれまでに 1 つの Web ページを複数の小ページに分割する Web ページ分割方式を提案している[2)]．これは，Web ページを構成する画像，ハイパーリン

† 株式会社 KDDI 研究所
KDDI R&D Laboratories Inc.

ク，テキスト等のコンテンツ間のつながりの強さを表すコンテンツ間距離をタグの構造の深さに着目して導出し，このコンテンツ間距離が固定的に決められた閾値を超える位置で分割する方式である．この方式は，HTML の文法を使用しないで HTML を扱うことが特徴であり，タグの省略等の非正則な HTML の場合でも問題なく処理が可能である利点を持つ．しかしながら，Web ページごとにコンテンツ間距離の大きさの分布が異なるため，すべての Web ページに適合する閾値を固定的に設定することが困難であるという課題があった．

そこで本稿では，Web ページのコンテンツ間距離の分布に基づき分割の閾値を動的に算出することで，すべての Web ページを適切に分割する Web ページ動的分割方式を提案する．さらに評価実験を行い提案方式の有効性を示す．

以降，2 章では，携帯電話を利用して Web ブラウジングを行うシステムの実現を目的とした関連研究とその課題から機能要件を導出する．3 章では，すでに提案しているコンテンツ間距離に基づく Web ページ分割方式の概要を述べる．4 章では，提案方式であるコンテンツ間距離の分布に基づく分割の閾値の決定する Web ページ動的分割方式を提案する．5 章では，評価実験とその結果，および Web ページ動的分割システムの実装例を紹介する．最後に，6 章でまとめを述べる．

2. 関連研究と機能要件の抽出

本稿では，携帯電話を利用して PC 向けの Web ページを効率的に閲覧可能にすることを目的とする．2.1 節では本方式と同様の目的を持つ関連研究とその課題を述べ，2.2 節では関連研究の課題から抽出した機能要件を示す．

2.1 関連研究と課題

関連研究は大きく 2 通りに分類できる．1 つは Web ページのレイアウトを携帯電話向けに変更する手法であり，もう 1 つは Web ページを再構築する手法である．以下にそれぞれの既存研究について述べる．

(1) Web ページのレイアウトを変更

携帯電話の狭い画面幅に合わせてブラウザが Web ページのレイアウトを縦長に変更する方式がある[4]．この方式は PC を利用した閲覧に近い画面表示を実現することを特徴とする．しかしながら，ユーザが必要とする情報は Web ページの一部のみである場合も多いことから[1),3)]，ユーザが見たい情報が下のほうに配置されていた場合には長いスクロール操作が必要となるため，探し出して閲覧するまでに手間がかかる課題がある．また，表形式の情報を携帯電話向けにレイアウト変更する方式がある[5),6)]．この方式は表の項目名を抽出して 1 行分ずつ見やすくレイアウトすることを特徴とする．しかしながらこれらの方式は表形式でレイアウトされていない情報には応用が困難である．これらの方式のほかにも，Web サイトのサイトマップのページを解析して携帯電話向けのメニューを自動生成する方式や[7),8)]，Web ページ中のリンクの重要度を利用者が判定し，重要度に基づくコンテンツのリストを自動生成する方式がある[9),10)]．これらの方式はサイトマップのページが必須であることや，選択したリンク先のページを閲覧するためには結局広い画面が必要となるという課題がある．

(2) Web ページを再構築

HTML の階層構造を解析して Web ページを小分割し，写真とその解説文のようにレイアウト上で関連性が高いコンテンツの集合（以下，オブジェクトと呼ぶ）を生成する方式がある[11),12)]．この方式では，大量の情報を持つ Web ページの場合でもオブジェクトごとに携帯電話の画面上に表示できるため，小さな画面でも容易に閲覧できるという特徴がある．しかしながら，HTML の構造を利用してオブジェクトの分割を行うため，正則な HTML でなければならない．しかしながら一般の Web ページは HTML の終了タグが省略されていることや，HTML には定義されていない不明なタグが挿入されていることも多いため，適用できない Web ページが多いという課題がある．またユーザのアクセス履歴等から嗜好情報を抽出し，単語の出現頻度をカウントしてユーザのプロファイルを作成することで，ユーザの嗜好に合わない部分を削除して小さな HTML を再構成する方式がある[13)]．この方式は，自動的に情報を取捨選択して HTML を小さくできる特徴を持つが，再構成した HTML が携帯電話で容易に閲覧可能な程度まで十分に小さくなるとは限らないという課題がある．また，1 つの Web ページを分割して複数の携帯電話の画面に割り当て，協力して閲覧することを目的とした方式がある[14)]．この方式は HTML 構造を重み付きの完全グラフに変換して分割することを特徴としているため，正しいグラフが生成できることが前提であり，非正則な HTML には適用できないという課題がある．

2.2 機 能 要 件

2.1 節で述べた関連研究の課題を集約すると，(1) 1 ページの情報量が多いと閲覧が不便であること，(2) 非正則な Web ページに対応できないこと，の

2 点となる．よって携帯電話向けの Web コンテンツ閲覧システムの機能要件として，次の 2 点があげられる．

（機能要件 1） 1 ページの情報量を小さくすること
携帯電話はユーザインタフェースの自由度が低いため，一度に表示する情報量を制限する必要がある．

（機能要件 2） あらゆる HTML に対応できること
閲覧の対象となる Web ページは，タグの省略や未知のタグの挿入等を含む非正則な Web ページに対しても処理できる必要がある．

3. コンテンツ間距離に基づく Web ページ分割方式

我々がすでに提案しているコンテンツ間距離に基づく Web ページ分割方式[2] について述べる．

機能要件 1 および 2 を満たすため，HTML を信号処理的に解析してコンテンツ間の距離を算出し，複数のコンテンツオブジェクトに分割する方式を提案する．提案方式は，タグの個数と相対的な深さのみに基づく方式でありタグの文法や種類によらないという特徴を持つ．そのため必要なタグの省略や未知タグの挿入等の非正則な HTML に対する耐性が高い方式であり，Chen らの方式[11] や前川らの方式[14] で検討されているような Web ページ全体のツリー構造が正しいことを前提として分割する位置を導出する方式とは異なる．

コンテンツ間距離を利用した Web ページの分割方式は次の (A)〜(C) の 3 つのステップからなる．ここで，ステップ (A) と (B) は SAX（Simple API for XML）パーサが出力した順に処理ができるため，高速な処理が可能である．なお本稿ではコンテンツを，（ア）HTML 中の `<a>` タグで指定されているアンカ，（イ）`<img>` タグで指定されている画像，（ウ）テキスト，の 3 種類と定義する．

(A) コンテンツの抽出
対象の Web ページの HTML ソースから 3 種類のコンテンツを抽出する．フォントタグ（`<font>`）やその他レイアウト用のタグ（`<table>`）等，コンテンツの修飾や配置のためのタグは携帯電話上で正しく再生できない場合が多いため，ここで削除する．ただし，`<table>` タグを利用した表形式の情報については配置が重要となるが，項目行を自動識別して表形式の情報のレイアウトを最適化する従来方式[5),6)] を利用すれば対応可能である．

(B) コンテンツ間距離の算出
隣接するコンテンツの間に挿入されているタグの数とその深さに基づき，コンテンツ間距離を算出する．ここでタグの深さとは，開始タグであれば直前のタグの深さに +1 し，終了タグがあれば −1 として部分的に累積して算出した値である．**図 1** に概念図を示す．横軸 x はタグの順序番号を表し，HTML ソースの先頭にある `<HTML>` タグから数えた番目とする．縦軸 y はタグの深さとする．コンテンツ a とコンテンツ b の距離を $S_{(a,b)}$ と定義し，式 (1) に従い導出する．式 (1) において，x_a および x_b はそれぞれコンテンツ a およびコンテンツ b のタグの順序番号であり，y_a および y_b はそれぞれコンテンツ a およびコンテンツ b のタグの深さである．また，$f_{ab}(i)$ はコンテンツ a，b 間における $x = i$ のときの y の値を示す．なお，式 (1) 中の x_a，x_b，y_a，y_b，$f_{ab}(i)$ は図 1 と対応しているので参照のこと．

$$S_{(a,b)} = \max\left\{ \sum_{i=x_a}^{x_b} \left|\max\{y_a, y_b\} - f_{ab}(i)\right|, \sum_{i=x_a}^{x_b} \left|\min\{y_a, y_b\} - f_{ab}(i)\right| \right\} \quad (1)$$

(C) コンテンツオブジェクトへの分割
コンテンツ間距離の大小を比較して 2 分割を再帰的に繰り返し，Web ページを複数のコンテンツオブジェクトに分割する．手順を以下に述べる．ここで，$N_1 > N_2$ とする．

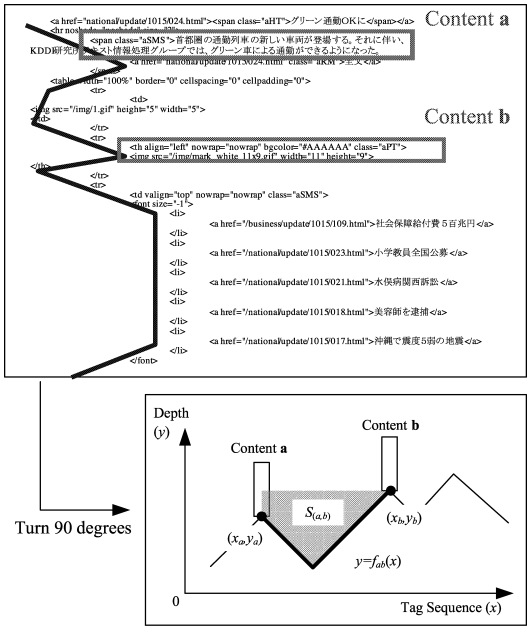

図 1 タグの深さとコンテンツ間距離
Fig. 1 Tag depth and content distance.

(1) Web ページ全体を 1 つのコンテンツオブジェクト（$ObjectID = \mathrm{root}$）とする．
(2) コンテンツオブジェクト内のコンテンツ間距離の最大値（$S_{\max}$）が，コンテンツオブジェクト内のコンテンツ間距離の平均値（S_{average}）の N_1 倍以上であれば，$S_{\max}$ の位置を分割点とする．
(3) (2) が真でない場合，S_{average} の N_2 倍以上かつ分割した場合のコンテンツ数の最小値が M 個以上であれば $S_{\max}$ の位置を分割点とする．
(4) 分割した場合は分割結果の左のコンテンツオブジェクトに移動し (2) に戻る．そうでなければ (5) に進む．
(5) 左のコンテンツオブジェクトであった場合は，右のコンテンツオブジェクトに移動し，(2) に戻る．
(6) 右のコンテンツオブジェクトであり，かつ $ObjectID \neq (root)$ の場合は，親コンテンツオブジェクトに移動し，(5) に戻る．それ以外は (7) に進む．
(7) 終了．

この方式では Web ページごとに最適となる分割の閾値 N_1，N_2 は Web ページごとに異なるため，特定の Web ページに最適な閾値を設定した場合でも，すべての Web ページに最適な値を設定するのは困難である．そのため，各々の Web ページに最適な閾値を動的に決定する方式が必要である．

4. 提案方式

3 章で述べた Web ページ分割方式の課題を解決するため，コンテンツ間距離の標準偏差に基づき Web ページごとに分割の閾値 N_1 と N_2 を動的に決定して Web ページを分割する方式を提案する．まず 4.1 節で提案方式の妥当性を検証するための予備実験を行った結果を示し，4.2 節で提案方式の詳細を述べる．

4.1 予備実験

コンテンツ間距離 S の分布に応じた N_1 と N_2 の設定方法を検討するため予備実験を行う．まず，Web ページごとの N_1 と N_2 の最適値を求め，N_1 と N_2 の相関について検証する．次に，N_1，N_2 とコンテンツ間距離とその統計値との関連について検証する．予備実験の詳細と結果を以下に述べる．

(1) N_1 と N_2 の相関について

Yahoo! カテゴリ[15] および Google ディレクトリ[16] から 106 のニュース提供ページと 104 の金融関連ページを任意に選択し，各 Web ページについて 3 章で記述した分割パラメータ N_1，N_2 を 1 から 16 の整数値の範囲で変化させてそれらの最適値を求めた．なお

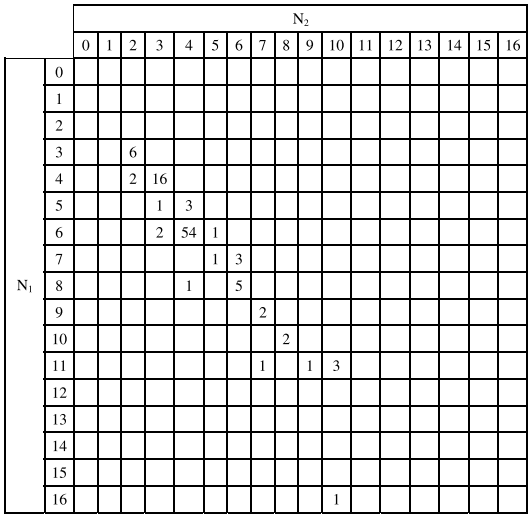

図 2 N_1，N_2 の最適値の分布（ニュース）
Fig. 2 Distribution of optimum value for N_1 and N_2 (news category).

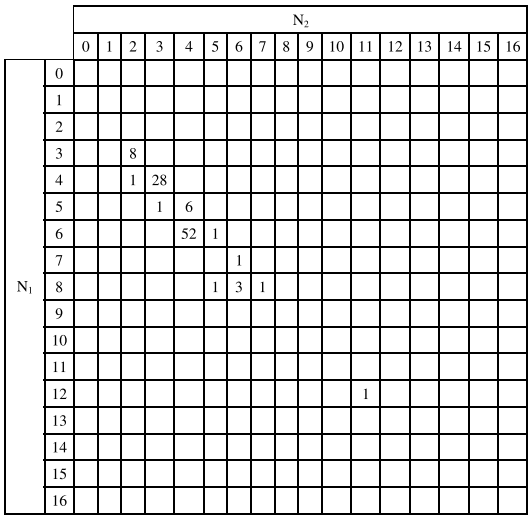

図 3 N_1，N_2 の最適値の分布（金融）
Fig. 3 Distribution of optimum value for N_1 and N_2 (finance category).

M は固定値 $M = 2$ とした．最適値は人の目で見て判断し，最も適切な位置でコンテンツオブジェクトに分割できた値とした．ニュース提供サイトと金融関連サイトの結果を**図 2** および**図 3** にそれぞれ示す．図 2 および図 3 において縦方向が N_1 の設定値，横方向が N_2 の設定値であり，該当した Web ページの数を各セルに記述した．

図 2，図 3 より，2 つの結果はほぼ同じ傾向であった．図 2 および図 3 における N_1 と N_2 のピアソンの積率相関係数を算出すると，それぞれ 0.943 と 0.938

表 1 $(6,4)$ と $(4,3)$ における Web ページの特徴比較
Table 1 Comparison of features in case of $(6,4)$ and $(4,3)$.

(N_1, N_2)	平均	標準偏差	コンテンツ数	HTML サイズ
$(6,4)$	9.5	23.7	1488.5	38.2
$(4,3)$	4.7	10.9	1026.4	36.7

平均：タグの深さの平均値
標準偏差：タグの深さの標準偏差の平均値
コンテンツ数：コンテンツ数の平均値 (個)
ファイルサイズ：HTML ファイルサイズの平均値 (Kbytes)

となった．このことから N_1 と N_2 は高い相関を持ち，ほぼ比例することが分かる．

(2) タグの深さの標準偏差と分割の閾値の関連

図 2 における $(N_1, N_2) = (6,4), (4,3)$ の場合について Web ページの特徴を比較した．結果を**表 1** に示す．

表 1 より，ファイルサイズがほぼ同等であるにもかかわらず，その他の数値はいずれも $N_1 = 6$ の方が大きな値となった．このことから (N_1, N_2) と平均，標準偏差，コンテンツ数の統計値に相関があると考えられる．

以上の予備実験結果より，Web サイトごとに適切な N_1 と N_2 を表 1 であげた統計値を利用して動的に決定可能であるといえる．ここで，N_1 と N_2 はコンテンツ間距離の差の大きさを判別する閾値であり，値のばらつきの度合いが影響することから，本稿ではこれらの統計値のうち意味が最も近い「標準偏差」に着目する．次節で閾値の動的決定方法について具体的に述べる．

4.2 コンテンツ間距離の分布に基づく閾値の動的決定方式

コンテンツ間距離の標準偏差に基づき，各 Web サイトごとに適切な N_1 と N_2 を動的に決定する方式を提案し，Web ページ動的分割アルゴリズムを示す．

4.2.1 N_1，N_2 の動的決定方式

任意の Web ページ t に適した閾値 N_{t1}，N_{t2} の算出手順を以下に示す．

（ア）基準値を設定する．

(1) コンテンツ間距離の分布の差異に基づき閾値 N_{t1}，N_{t2} を決定するため，基準となる Web ページ（以下，基準ページと呼ぶ）を任意に選択する．

(2) 基準ページを最適に分割可能な N_1，N_2 を実験的に決定し，これを N_{b1}，N_{b2} とする．

(3) 基準ページのコンテンツ間距離 $S_{b(i,i+1)}$ の集合を 3 章の手順 (C) に従い導出する．

(4) 式 (2) に従い，σ_{S_b} を算出する．

$$\sigma_{S_b} = \sqrt{\frac{\sum_{i=1}^{n_b-1}\left(\overline{S}_b - S_{b(i,i+1)}\right)^2}{n_b - 1}} \quad (2)$$

$\overline{S}_b$：	基準ページのコンテンツ間距離の平均値
$S_{b(i,i+1)}$：	基準ページのコンテンツ i とコンテンツ $i+1$ 間の距離
n_b：	基準ページで抽出したコンテンツ数

（イ）対象の Web ページ用 N_{t1}，N_{t2} を算出する．

(1) 対象ページのコンテンツ間距離 $S_{t(i,i+1)}$ の集合を 3 章の手順 (C) に従い導出する．

(2) 式 (3) に従い，標準偏差 σ_{S_t} を算出する．

$$\sigma_{S_t} = \sqrt{\frac{\sum_{i=1}^{n-1}\left(\overline{S}_t - S_{t(i,i+1)}\right)^2}{n_t - 1}} \quad (3)$$

(3) 式 (4) および式 (5) に従い，N_{t1}，N_{t2} を算出する．ここで α は正の実数値（$\alpha > 0$）とし，$t = b$ のとき $N_{t1} = N_{b1}$，$N_{t2} = N_{b2}$ とする．

$$N_{t1} = N_{b1} + N_{b1} * \left(\frac{\sigma_{S_t}}{\sigma_{S_b}} - 1\right) * \alpha \quad (4)$$

$$N_{t2} = N_{b2} + N_{b2} * \left(\frac{\sigma_{S_t}}{\sigma_{S_b}} - 1\right) * \alpha \quad (5)$$

$\overline{S_t}$：	分割対象の Web ページの距離の平均値
$S_{t(i,i+1)}$：	分割対象の Web ページのコンテンツ i とコンテンツ $i+1$ 間の距離
n_t：	分割対象の Web ページで抽出したコンテンツ数

4.2.2 閾値の動的決定機能を有する分割アルゴリズム

提案した閾値の動的決定方式を組み込んだ Web ページ分割アルゴリズムを**図 4** に示す．基本的には 3 章の手順と同様であるが，N_{t1}，N_{t2} の導出手順が加わっている．

5. 評　　価

提案方式の有効性を確認するため，評価実験を行った．5.1 節で実験内容について述べ，5.2 節で結果について述べる．5.3 節で本システムの実装例と動作例を示す．

5.1 実 験 条 件

(1) 比較対象の方式

方式 A　論文に基づき独自に作成したパーサを利用して再現した Chen らの方式[11]

方式 B-1　**表 2** のリスト外の特定のページ[17] に

```
[Segmentation Algorithm]
objectID = (root); x = LEFT;
D_{objectID,x} = (Whole of contents);
  σ_{S_t} = Sdeviation{Standard deviation of content distances};
  N_{t1} = N_{b1} + N_{b1} * (σ_{S_t}/σ_{S_b} − 1) * α;
  N_{t2} = N_{b2} + N_{b2} * (σ_{S_t}/σ_{S_b} − 1) * α;
Segment(D_{objectID,x}){
  S_max = max{S_{t(i,i+1)} : i ∈ D_{objectID,x}};
  S_average = average{S_{t(i,i+1)} : i ∈ D_{objectID,x}};
  C_segmented = (Minimum number of tags in a group
                 when it is segmented)
  if(S_max > N_{t1} * S_average){
    (Segment at the S_max);
    ChildObjectID = (objectID of created object by segmentation);
    Segment(D_{ChildObjectID,LEFT});
  };
  else if(S_max > N_{t2} * S_average & C_segmented > M){
    (Segment at the S_max);
    ChildObjectID = (objectID of created object by segmentation);
    Segment(D_{ChildObjectID,LEFT});
  };
  if(x = LEFT){
    Segment(D_{objectID,RIGHT});
    Return;
  };
  else if(x == RIGHT & objectID ≠ (root)){
    Return;
  };
  else{
    END;
  };
};
```

図 4 N_{t1}，N_{t2} の動的算出機能を有する分割アルゴリズム
Fig. 4 Segmentation algorithm with dynamic calculation of N_{t1} and N_{t2}.

対して適合率が最大化するように最適化した閾値を固定的に設定した Web ページ分割方式[2)]

方式 B-2 表 2 にリストした Web ページ対して適合率が最大になるように最適化した閾値を固定的に設定した Web ページ分割方式[2)]

方式 B-3 表 2 にリストした個々の Web ページの F 値が最大になるように最適化した閾値を個々に設定した Web ページ分割方式[2)]

提案方式 表 2 のリスト外の特定のページ[17)] を基準ページとして，閾値を動的に決定する Web ページ分割方式

(2) 評価対象の Web サイト

Chen ら[11)] が評価対象とした Web サイトのうち，現在アクセス可能な 37 の Web サイトの各トップページを評価対象とする．Web サイト一覧を表 2 に示す．

(3) 評価方法

各方式を PC 上に実装し，表 2 に示した Web サイトから HTML を取得して各方式で分割処理を行った．各方式の分割結果について適合率と再現率を式 (6) および式 (7) に従い算出し，さらに F 値を算出して対象サイトの平均値を比較した．ここで「正解」とは，各方式で自動的に判定した個々の分割位置が，PC のブラウザで対象の Web サイトを見た場合の評価者の主観により最適と判定した分割位置のいずれかと合致することを指す．

表 2 実験対象とする Web サイト
Table 2 Web sites for using in evaluation.

JobsOnline.com	Yahoo.com
flowgo.com	Msn.com
Americangreetings.com	Aol.com
Mypoints.net	Microsoft.com
Cnn.com	Altavista Search Services
Bizrate.com	Go.com
Mapquest.com	Amazon.com
Weather.com	Nbci.com
Infospace.com	Ebay.com
Iwin.com	Bluemountain.com
Espn.com	Lycos.com
Colonize.com	Looksmart.com
Travelocity.com	Cnet.com
Windowsmedia.com	Angelfire.com
Ivillage.com	Tripod.com
Disney Online	Iwon.com
Zmedia.com	Zdnet.com
Google.com	Msnbc.com
Earthlink.net	

表 3 パラメータ設定値
Table 3 Parameter settings.

方式	パラメータと値
方式 B-1	$N_1 = 2.6$，$N_2 = 1.7$，$M = 2$
方式 B-2	$N_1 = 2.9$，$N_2 = 2.6$，$M = 2$
方式 B-3	N_1，N_2 は各 Web ページの最適値，$M = 2$
提案方式	$N_{b_1} = 3.4$，$N_{b_2} = 2.3$，$M = 2$，$\alpha = 0.36$

方式 B-1，方式 B-2，方式 B-3，および提案方式の各パラメータの設定値を**表 3** に示す．

$$\text{適合率} = \frac{(a)\ \text{正解した分割位置数}}{(b)\ \text{全分割位置数}} \tag{6}$$

$$\text{再現率} = \frac{(a)\ \text{正解した分割位置数}}{(c)\ \text{全正解分割位置数}} \tag{7}$$

(a) 正解した分割位置数：	各方式の分割位置のうち正解した数
(b) 全分割位置数：	各方式で分割した位置数
(c) 全正解分割位置数：	人手で判断した正解位置数

5.2 実験結果

（ア）分割性能比較

実験結果を**表 4** に示す．表 4 より方式 B-1 では F 値が 0.59 となり，方式 A の 0.45 よりも高い結果となった．ただし，方式 A の結果については我々が Chen らの方式の論文[11)] から読み取れる情報に基づき実装しているため，Chen らの実装とは異なる可能性もあることから参考値とする．閾値を最適化した方式 B-2 では F 値が 0.58 となり方式 B とほとんど違いはみられなかった．実験では適合率を最大化するようにパラメータ調整しているため適合率で比較すると，方式

表 4 実験結果
Table 4 Evaluation results.

方式	適合率	再現率	F 値
方式 A	0.71	0.33	0.45
方式 B-1	0.64	0.55	0.59
方式 B-2	0.80	0.45	0.58
方式 B-3	0.72	0.61	0.66
提案方式	0.87	0.50	0.64

B-1 が 0.64 であるのに対し，方式 B-2 は 0.80 と大幅に上回っていることが分かる．また，提案方式では F 値が 0.65 となり方式 B-2 よりもさらに高い値が得られた．また，方式 B-3 の結果は提案方式の理想値に等しい意味を持つが，提案方式の結果はその理想値に近い値が得られた．以上の結果から，コンテンツ間距離の標準偏差に基づく閾値の動的決定が有効に作用していることを示し，提案方式の有効性を確認できた．

（イ）考察

提案方式では，対象とした表 2 の Web ページの中で，aol.com の Web ページの再現率が最も低い結果となった．18 の正解位置数に対して 1 カ所しか分割位置を検出できていないことが原因であり，再現率は 0.06 となっていた．この理由として，提案方式の特徴はレイアウト用等の直接表示されないタグに着目して分割処理を行う点であることから，それらのタグがない，あるいは少ない場合には分割が難しい結果となると考えられる．実際に，aol.com のページの HTML ソースを分析すると，コンテンツのレイアウトをほぼ完全に CSS（Cascading Style Sheets）ファイルで記述しており，HTML ソース自体は非常にシンプルに記述されていた．このとき提案方式はコンテンツ間距離をほとんど 0 と算出してしまうことから，分割する点をうまく抽出することができなかった．

5.3 システム実装例

提案方式の実装例として携帯電話向け Web 閲覧システムを**図 5** に示す．本システムは，(1) ユーザからの要求を受付ける Request Reception，(2) 要求に応じて対象 Web ページの HTML を取得する Browser Emulator，(3) 取得した HTML を提案方式により分割する Segmentation Processor，(4) 分割処理結果を再構成して新たな HTML を生成する HTML Re-builder の 4 つから構成される．

図 6 に示した写真つきのニュースを提供している Web ページ[17] に対して本システムを適用すると，たとえば破線で囲まれた写真とテキストの組をコンテンツオブジェクトとして正しく認識し，**図 7** に示すように携帯電話で閲覧することができた．

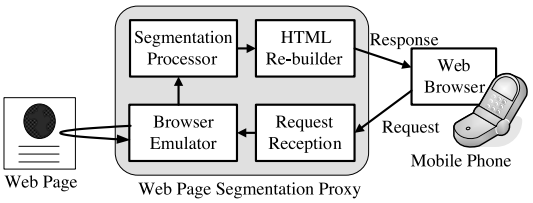

図 5 携帯電話向け Web 情報閲覧システム
Fig. 5 Web information display system for mobile phone.

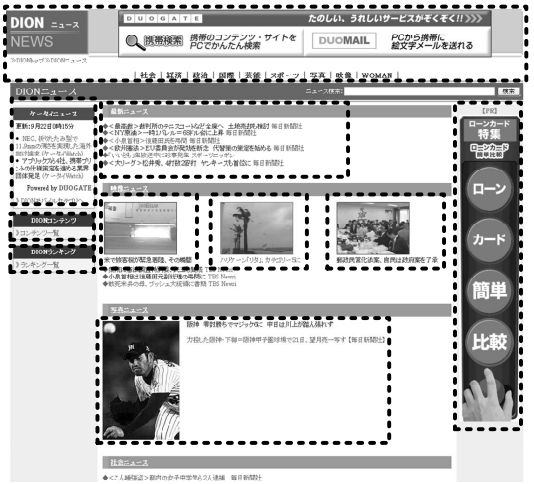

図 6 Web ページ例
Fig. 6 Example of Web page.

図 7 コンテンツオブジェクトのリスト（左）とコンテンツオブジェクトの表示（右）
Fig. 7 List and display of contents objects.

6. おわりに

本稿では，PC 向けの Web ページを小分割して HTML を再構成し，画面の小さい携帯電話でも容易に閲覧するためのコンテンツ間距離に基づく Web ページ自動分割方式について検討した．これまでの方法ではすべての Web ページに最適となる分割閾値を固定的に設定することが困難であったため，Web ページごとに適切な分割閾値をコンテンツ間距離の標準偏差

に基づき動的に決定する方式を提案した．方式の妥当性を予備実験により示し，また分割精度について従来方式と比較する評価実験を行った．評価実験ではF値が0.64で従来方式よりも高い値が得られ，提案方式の有効性を示した．

謝辞 日頃ご指導いただくKDDI研究所浅見代表取締役所長，および中島執行役員に深く感謝いたします．

参 考 文 献

1) 服部 元，松本一則，菅谷史昭：表形式情報集約のための連想性の高いオブジェクトラベルの自動抽出方式，*3rd Joint Agent Workshops & Symposium* (2004).
2) 服部 元，松本一則，菅谷史昭：タグの深さを利用したコンテンツ間距離に基づくWebページの自動分割方式，日本データベース学会Letters, Vol.4, No.1 (2005).
3) 山田誠二，中井有紀：対話的分類学習によるWebページの部分更新モニタリング，人工知能学会論文誌，Vol.17, No.5 (2002).
4) Small Screen Rendering (Opera Software ASA). http://www.opera.com/products/mobile/smallscreen/
5) Chen, Y., Ma, W.-Y. and Zhang, H.-J.: Improving Web Browsing on Small Devices Based on Table Classification, *12th International World Wide Web Conference*, 20–24 (May 2003).
6) 増田英孝，塚本修一，安富大輔，中川裕志：HTMLの表形式データの構造認識と携帯端末表示への応用，情報処理学会論文誌：データベース，Vol.44, No.12 (2003).
7) Buchanan, G., Farrant, S., Jones, M. and Thimbleby, H.: Improving Mobile Internet Usability, *Proc. 10th International World Wide Web Conference*, Hong Kong, China (2001).
8) Jones, M., Buchanan, G. and Thimbleby, H., Sorting out Searching on Small Screen Devices, *Conference on Mobile HCI* (2002).
9) Buyukkokten, O., Garcia-Molina, H. and Paepcke, A.: Seeing the whole in parts: Text summarization for web browsing on handheld devices, *Proc. 10th International World Wide Web Conference* (2001).
10) Buyukkokten, O., Garcia-Molina, H., Paepcke, A. and Winograd, T.: Power browser: Efficient web browsing for PDAs, *Proc. Human-Computer Interaction Conference 2000* (2000).
11) Chen, Y., Ma, W. and Zhang, H.: Detecting web page structure for adaptive viewing on small form factor devices, *Proc. World Wide Web Conference 2003* (2003).
12) 前川卓也，原 隆浩，西尾章治郎：複数のモバイルユーザのためのWebページ分割を用いた協調Webブラウジングシステム，情報処理学会モバイルコンピューティングとユビキタス通信研究会，Vol.2004, No.114 (MBL 31) (2004).
13) Anderson, C.R., Domingos, P. and Weld, D.S.: Personalizing web sites for mobile users, *Proc. 10th International World Wide Web Conference* (2001).
14) 前川卓也，上向俊晃，原 隆浩，西尾章治郎：複数のモバイル端末による協調ブラウジングのための木構造型コンテンツ記述方式と分割方式，情報処理学会論文誌：データベース，Vol.45, No.SIG7 (TOD 22) (2004).
15) Yahoo! Japan カテゴリ. http://www.yahoo.co.jp/
16) Google ディレクトリ. http://www.google.co.jp/
17) DION ニュースサイト. http://newstopics.dion.ne.jp/pubnews/
18) 社団法人電気通信事業者協会. http://www.tca.or.jp/index.html

(平成17年9月19日受付)
(平成18年3月13日採録)

(担当編集委員 石川 博，有次 正義，片山 薫，木俵 豊，土田 正士)

服部 元（正会員）

平成8年神戸大学工学部電気電子工学科卒業．平成10年同大学大学院修士課程修了．同年国際電信電話（株）（現KDDI（株））入社．現在，（株）KDDI研究所テキスト情報処理グループ研究員．この間，ネットワーク管理，高度交通システム，ソフトウェアエージェント，Webアプリケーションの研究開発に従事．平成15年電子情報通信学会学術奨励賞受賞．電子情報通信学会，日本データベース学会各会員．

松本 一則（正会員）

昭和 59 年京都大学工学部情報工学科卒業．昭和 61 年同大学大学院修士課程修了．同年国際電信電話（株）（現 KDDI（株））入社．現在，（株）KDDI 研究所テキスト情報処理グループ主任研究員．この間，マルチメディア検索，コンテンツ配信の研究開発に従事．平成 10 年人工知能学会研究奨励賞，平成 12 年度電子情報通信学会論文賞を各受賞．電子情報通信学会会員．

菅谷 史昭（正会員）

昭和 57 東北大学工学部通信工学科卒業．昭和 59 年同大学大学院修士課程修了．同年国際電信電話（株）（現 KDDI（株））入社．平成 9〜14 年まで ATR 音声翻訳通信研究所に出向．平成 14 年 KDDI（株）復帰．現在，（株）KDDI 研究所テキスト情報処理グループリーダ．この間，情報検索，e-Learning，音声翻訳評価の研究開発に従事．平成 3 年電子情報通信学会学術奨励賞受賞．電子情報通信学会，日本音響学会各会員．工学博士．