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

Auto Web Page Distilling Scheme Based on Content Distance Using Relative Tag Hierarchy

服部 元[†1] 松本 一則[†2]
菅谷 史昭[†3]

Gen HATTORI Kazunori MATSUMOTO
Fumiaki SUGAYA

携帯端末を利用してインターネット上の一般の Web ページを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いるため，携帯端末でそれらの情報を容易に閲覧するためには，Web ページを階層的に再構築する等の方法が必要である．既存の方式として，HTML の厳密なタグ構造を解析して Web ページを分割・再構築する方式が提案されているが，タグの省略等を含む非正則な HTML には対応できない等の課題がある．そこで本稿では，Web ページを部分的に解析して得られるタグの数や深さ等の HTML タグの相対的な構造を利用して算出したコンテンツ間距離に基づき，大小 2 種類の閾値を利用して Web ページの分割点を導出する方式を提案した．評価実験を行い非正則な Web ページが数多く存在することを示し，また提案方式が約半数の Web ページを適切に分割できることを示した．また分割性能を向上する方法について検討し，タグの統計的性質を利用した性能向上の見通しが得られた．

The demand of retrieving information from general Web pages using a mobile terminal is increasing. However, most of general Web pages in the Internet are designed for PC users. In order to read these information easily with a mobile terminal, the methods which reconstructs the pages hierarchically, and so on, are required. Therefore, we have to divide a Web page into small articles and rebuild stratificational structure of Web pages. Although there are conventional scheme, which divides a Web page small, the scheme had problem that they cannot adapt to the Web pages to which one or more tags were abbreviated. In this paper, we propose a new automatic Web page distilling scheme by using the distance between contents based on the relative HTML tag hierarchy, which is the number and depth of HTML tags in a Web page. We conducted the evaluation experiment and show that many un-regular Web pages exist, and our proposed scheme can divide appropriately a half number of Web pages. Moreover, we examined how to improve the accuracy of distilling, and acquired the prospect of the improvement in accuracy using the statistical property of tags.

†1 正会員 株式会社KDDI研究所 gen@kddilabs.jp
†2 非会員 株式会社KDDI研究所 matsu@kddilabs.jp
†3 非会員 株式会社KDDI研究所 fsugaya@kddilabs.jp

1. はじめに

近年，携帯電話やPDA等の携帯端末の普及が進んでいる．これらの多くはWebにアクセスする機能を持っており，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の収集が可能である．一方，携帯端末向けのWebページの数はPC向けと比較してはるかに少なく，また携帯端末は狭小な画面と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，携帯端末向けのWebページが持つ情報量は少なく調整される傾向にある．そのため，携帯端末から情報量が豊富な一般のWebページを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，画面のレイアウトや情報量をPCユーザ向けに調整して作成されており，そのままでは携帯端末での閲覧はできない．またこのときユーザが関心を持つ情報はWebページの一部のみである場合も多いことから[1]，Webページの情報を階層化する等，ユーザが必要とする情報を容易に選択可能とすることが重要である．既存の方式としてHTML (Hyper Text Markup Language)の文書構造を全体的に解析して得られる絶対的なタグの構造を利用し，Webページを小分割して情報を階層化する方式[2]が提案されているが，HTMLタグの省略等を含む非正則なWebページには適用できない課題があった．

そこで本稿では，HTMLの部分的な解析により得られるタグの数や深さ等の相対的な階層構造を利用して算出した距離に基づき，Webページを小分割する方式を提案する．本方式は部分的かつ相対的な階層構造を利用することからHTMLの非正則な記述が解析の成否に影響しないため，非正則なWebページにも対応できる．また評価実験を行い，本方式の有効性の検証を行う．

2. 従来方式の課題と機能要件

本稿の目的は，ユーザが携帯端末を利用して，PC 向けの Web ページ上の情報から必要な情報を容易に選択して閲覧するためのシステムを実現することとする．ここで Web ページの写真やテキスト，ハイパーリンク等の可視的な情報をコンテンツと呼び，またコンテンツの小規模な集合をコンテンツオブジェクトと呼ぶ．

従来方式として，端末の画面の幅に合わせて Web ページのレイアウトを変更する方式がある[3]．ただしこの方式では縦長の Web ページにレイアウトするのみで情報の階層化は実現していないため，ユーザが必要とする情報にたどり着くまでに時間を要する場合もあり，容易な情報の選択を実現しているとはいえない．他の方式として，DOM (Document Object Model)パーサを利用して対象となるWebページ全体のタグを分析し，得られた絶対的なタグの階層構造に基づきコンテンツオブジェクトに分割する方式がある[2]．コンテンツオブジェクトをユーザが取捨選択できるように情報を階層化することで，携帯端末のような小さい画面でも容易に PC 向けの Web ページを閲覧できる環境を実現している．しかしながら，タグの省略や未知のタグの挿入等を含む非正則な HTML には対応できない問題がある．

以上の検討から，PC向けのWebページを携帯端末で閲覧するシステムの機能要件として，次の2つが挙げられる．いずれの従来方式も，2つの機能要件を同時には満たしていない．

(1)1ページの情報量を小さくすること
PC向けのWebページは情報量が豊富であるが携帯電話のユーザインタフェースが貧弱で自由度が低い．そのため，一度に表示する情報量を制限する必要がある．

(2)あらゆるHTMLに対応できること
HTMLはXML (Extended Markup Language)のような厳密な記述規則はないため非正則なHTMLが多く存在する．そのため，非正則なHTMLであっても対応する必要がある．

3. 提案方式

2 つの機能要件を満足する Web ページの分割･階層化方式として，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を提案する．機能要件(1)を満たすため，Web ページを小分割する方針とし，機能要件(2)を満たすため，HTML を部分的に解析して得られる相対的なタグ構造を利用する方針とする．

3.1. 概要

提案方式の概要について述べる．なお提案方式ではコンテンツを，(ア)HTML 中の<a>タグで指定されているアンカー，(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．

図 1 に示す Web ページは破線で示した 3 つのコンテンツオブジェクトからなる．HTML のソースを図 2 に示す．提案方式はコンテンツ間の距離が大きい部分を分割点としてコンテンツオブジェクトに分割する．例えば図 2 ではコンテンツオブジェクト間の距離として示している部分の距離が大きいためここを分割点として判断する．判断結果に基づきコンテンツオブジェクトに分割して HTML を再構成し，携帯端末で表示する．表示例を図 3 に示す．

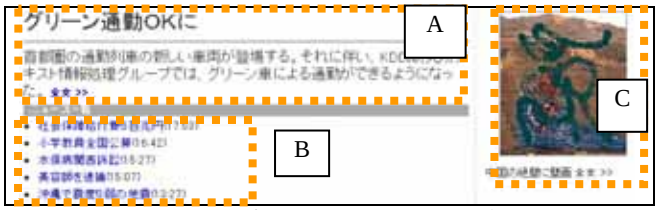

図 1 対象とする Web ページの例
Fig. 1 Example of Web page

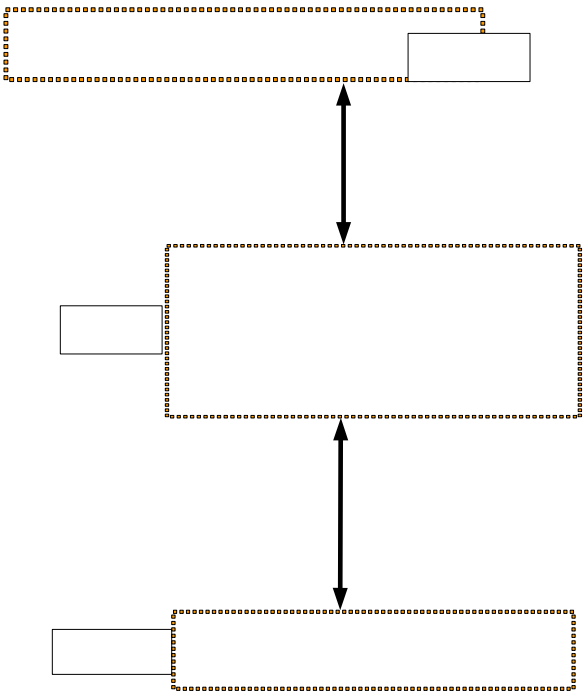

図 2 コンテンツ間距離
Fig. 2 Content Distances

図 3 携帯端末の画面例
Fig. 3 Display Example of Mobile Terminal

3.2. 相対的なタグ構造を利用したコンテンツ間距離の算出方式と Web ページ分割方式

提案方式は次の(A)コンテンツの抽出，(B)コンテンツ間距離の算出，(C)コンテンツオブジェクトへの分割，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)パーサの出力順に処理できるため，高速な処理が可能である．各ステップの詳細を以下に述べる．

(A) コンテンツの抽出

HTML の先頭から順にタグを検索し，HTML 中の<a>タグで指定されているアンカー，<img>タグで指定されている画像，およびテキストを抽出する．

(ア) アンカーの抽出

開始タグ<a>と終了タグで囲まれた部分を 1 コンテンツとする．<img>タグで指定した画像とテキストを含む場合はまとめて 1 つのコンテンツとする．

(イ) 画像の抽出

(ア)に含まれるものを除く画像をコンテンツとする．ただし，<img>タグは本来の画像として表示する以外にも，小さなスペースを作るための利用や記事のセパレータとしての利用が多い．このような可視的でない画像をコンテンツと見なさないため，<img>タグの画像をコンテンツとするための条件を以下の「(a)または(b)」とする．

(a)画像のサイズが縦P_c (pixels)，かつ横P_r (pixels)以上であること

(b)<img>タグの属性値”alt”で指定された代替テキストが記述されていること

(ウ) テキストの抽出

(ア)に含まれるテキストを除き，タグに挟まれているテキストをコンテンツとする．

(B) コンテンツ間距離の算出

コンテンツ間距離を算出する．コンテンツ間距離とは HTML ソース上で隣接するコンテンツの近接度を表す．言語を限定する場合は自然言語処理によりコンテンツ間の近接度を求める方法も考えられることができるが，WWW は文字通り World Wide な環境であるため言語に依存しない方法が望ましい．

PC向けのWebページは<table>等のHTMLのタグを利用して見た目のレイアウトを制御しており，近接しないコンテンツ間ではHTMLの構造が大きく変化している．このような部分はタグの深さが大きく変化する部分であることから，(1)コンテンツ間にあるタグの数と(2)それらのタグの深さ変化の度合いが，コンテンツ間距離の指標となるといえる．そこで提案方式では，図 4 に示すようにコンテンツ間にあるタグとその深さ，およびコンテンツの深さで囲まれた部分の面積を算出し，これをコンテンツ間距離と定義する．

コンテンツA, B間の距離S_{ab} は図 5 に示すフローにより算出する. ここで, $y= f(x)$ はタグの出現順序(x) に対する深さ(y)を表す. また, S_aとS_bを算出してその最大値を選択しているのは, 図 4 に示すような$y = f(x)$が谷型の場合だけでなく, 山型の場合についても対応するためである.

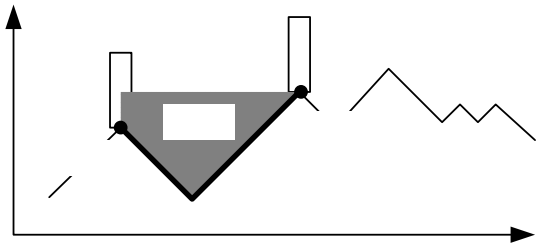

図 4 タグの深さとコンテンツ間距離
Fig. 4 Tag Depth and Content Distance

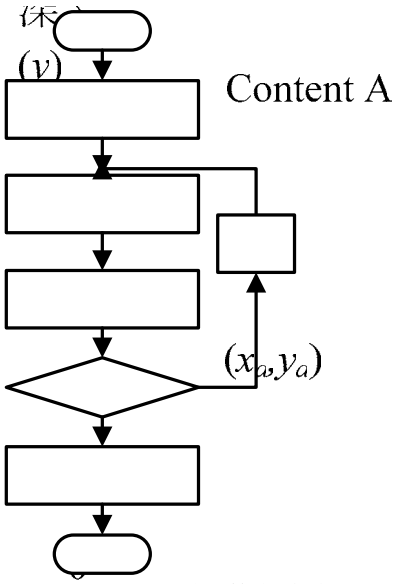

図 5 コンテンツ間距離の算出手順
Fig. 5 Calculation of Content Distance

(C) コンテンツオブジェクトへの分割

算出したコンテンツ間距離に基づき分割点を決定し, Web ページをコンテンツオブジェクトに分割する. 手順を以下に示す.

(1) Web ページ全体を 1 つのコンテンツオブジェクト (ObjectID =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N_3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側のコンテンツオブジェクトに移動し(2)へ. 分割しなかった場合は(5)へ.
(5) 左側のコンテンツオブジェクトであった場合は, 右側のコンテンツオブジェクトに移動し, (2)へ.
(6) 右側のコンテンツオブジェクトで, かつ ObjectID<>(root) の場合は, 親コンテンツオブジェクトに移動し, (5)へ.
(7) 終了.

4. 評価

4.1 非正則な HTML の割合

予備実験として, W3C の Web サイト[4]と, 文献[2]の評価実験で使用している Web ページのうち現時点でアクセス可能な Web ページ(39 件)に対し, Apache XML Project が開発している DOM パーサ(Xerces)[5]でツリーの構築を試みた. その結果, W3C を除く全てのページでツリー構築エラーとなり非正則な HTML であると判明した. 非正則な HTML のパターンには以下のような場合があった.

(1) 終了タグが必要であるにも関わらず記述していない
(2) タグの入れ子構造がクロスしている
(3) タグ名のスペルミス, 不明なタグ

但し, 処理時間が余分にかかる問題はあるが, ある程度の不足やミスであれば HTML の修復ツールを利用して自動的に修復可能である. そこで HTML Tidy Library Project が開発した HTML の修復ツールである TIDY[6]を利用した場合について, TIDY でも修復不能であった致命的なエラーのある HTML の数を検証した. 対象とした Web ページは, Yahoo!Japan のカテゴリ情報から, ニュース, 料理, 自然, エンタテイメント, スポーツのカテゴリに属する Web ページからそれぞれ 70 ページ以上を任意に選択したものとした. 結果を表 1 に示す.

表 1 エラーのある Web ページ数と割合
Table 1 Number and Ratio of Error Web Pages

カテゴリ	Web ページ数	エラー数	割合(%)
ニュース	106	9	8.5
料理	70	19	27.1
自然	71	14	20.0
エンタテイメント	71	7	9.9
スポーツ	106	19	18.0

ニュース情報を提供している Web ページは 8.5%程度であり, エンタテイメントのカテゴリでは約 10%であった. これに対し, 料理や自然のカテゴリではそれぞれ 27.1%, 20.0%となり, エラーのある Web ページの割合がかなり高い結果となった. カテゴリによりばらつきはあるが 10%～30%程度は含まれていると考えられる. これらの Web ページに対しては HTML 修復ツールを利用しても従来方式ではツリーを正しく作成することはできないが, 提案方式は対応可能である.

4.2 有効性の評価

提案方式を実装し有効性の評価を行う. ニュースを提供している 106 のサイトを対象とし, 各Webサイトについて 3.2 節(C)で記述した分割パラメタN_1およびN_2を変化させてそれらの最適値を求めた. なおN_3 = 2 とし, P_c, P_rは 100 とした. 最適値は人の目で見て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. 結果を図 6 に示す. 図 6 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, 該当するWebページ数を各セルに記述した. N_1= 6, N_2 = 4 の場合が最も多い 54 となり, 全 106 サイトのうち約 51%を占めた. よって, 提案方式は対象としたニュースサイトに対しては, N_1= 6, N_2= 4 と設定することにより約 51%のWebサイトに対して最適な分割が可能であったといえる.

N_1 \ N_2	0	1	2	3	4	5	6	7	8	9	10	11
0												
1												
2												
3			6									
4			2	16								
5				1	3							
6				2	54	1						
7						1	3					
8					1		5					
9								2				
10									2			
11								1		1	3	
12												
13												
14												
15												
16											1	

図 6 N_1とN_2の最適値の分布
Fig. 6 Distribution of Optimum Value for N_1 and N_2

4.3 分割パラメタの動的設定方法

より多くのWebページに適応可能とするため, パラメタを動的に設定する方法について検討する. 例えば図 6 において 2 番目に多いのはN_1 = 4, N_2 = 3の場合の16であり, 全体の約15%を占める. 例えばこの領域に属するWebページにも対応することができれば合計で約 66%に対応可能となる. そこでN_1 = 6 とN_1 = 4 の場合について, 属するWebページの特徴を以下の 4 つの項目について比較した. 結果を表 2 に示す.

- 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)

表 2 N_1= 6 とN_1 = 4 におけるWebページの特徴比較
Table 2 Comparison of Features of N_1 = 6 and N_1 = 4

	平均	標準偏差	コンテンツ数	ファイルサイズ
N_1=6	9.5	23.7	1488.5	38.2
N_1=4	4.7	10.9	1026.4	36.7

表 2 より, ファイルサイズがほぼ同等であるにも関わらず, いずれの数値もN_1 = 6 の方が大きな値となった. また図 6 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 0.94 となった. 強い相関が確認でき, 片方が決まればもう一方も容易に決定できるといえる. 以上の検討より, これらの値を利用してN_1およびN_2を自動的に決定可能であると考えられる.

5. まとめと今後の課題

本稿では, PC 向けの Web ページを小分割して HTML を再構成し, 画面の小さい携帯端末でも容易に閲覧するための新たな方式を提案した. 提案方式は HTML を部分的に解析して得られる相対的なタグの階層構造を利用してコンテンツ間距離を算出し, その大小を区切りのパラメタとして小分割することが特徴である. これにより従来方式では対応できなかった非正則な HTML にも対応可能とした. 評価実験を行い, 非正則な HTML がカテゴリに応じて 10%～30%程度存在することを示し, 提案方式が約 51% の Web サイトに対して最適な分割が可能であることを示した. さらに Web ページのいくつかの統計的な特徴から分割のパラメタを推定する方法により, 分割可能な Web ページの割合を増やすことができる見通しを得た. 今後の課題として, 分割パラメタの推定方式の評価ならびに既存の方式との性能比較を行うことが挙げられる.

[謝辞]

日頃ご指導頂く 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致します.

[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] Y. Chen, W. Ma, and H. Zhang, ”Detecting Web page structure for adaptive viewing on small form factor devices,” in Proc. World Wide Web Conference 2003, 2003.
[3] Small Screen Rendering (Opera Software ASA), http://www.opera.com/products/mobile/smallscreen/.
[4] World Wide Web Consortium, http://www.w3.org/.
[5] Apache XML project Xerces2 Java Parser, http://xml.apache.org/xerces2-j/.
[6] HTML Tidy Library Project, http://tidy.sourceforge.net/.

服部 元 Gen HATTORI
平成 8 年神戸大・工・電気電子工学卒業. 平成 10 年同大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, ITS, ソフトウェアエージェントの研究開発に従事. 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, 情報処理学会, 日本データベース学会各会員.

松本 一則 Kazunori MATSUMOTO
昭和 59 年京都大・工・情報工学卒業. 昭和 61 年同大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, 情報処理学会各会員.

菅谷 史昭 Fumiaki SUGAYA
昭和 57 東北大・工・通信工学卒業. 昭和 59 年同大大学院修士課程了. 同年国際電信電話(株)(現 KDDI(株))入社. 平成 9 年より平成 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, 情報処理学会各会員. 工博.