類似ページと差分をもちいたスマートフォン用 Webページ変換方法の提案

2009SE114 河村祥希　2009SE130 小島立士　2009SE167 溝口亮太
指導教員：蜂巣吉成

1　はじめに

近年，日本におけるスマートフォンの利用が増加し，スマートフォンでの Web ページ閲覧が頻繁に行なわれるようになった．スマートフォンでの閲覧を考えて作られていない Web ページでは，スクロールやページの拡大・縮小を行う必要があり，利用者が不便に感じることが多くなってきている [1]．スマートフォンでの閲覧に適した Web ページが増えつつあるが，対応していない Web ページは現在も多く存在する．企業サイト，EC サイトに関してはスマートフォン閲覧に対応しているサイトは全体の半分弱とする調査結果もある [2].

Web ページの自動変換は存在するが開発者向けであり，利用者が独自で見やすくなるように Web ページを変換することは，専門の知識がなければ容易ではない．既存の自動変換ツールなどを用いて Web ページを変換したとしても利用者が望むようなレイアウトにならない場合がある．Web ページの閲覧者は，Web サイト作成者がスマートフォン用サイトに対応しない限り，閲覧し易さに変化はない．パーソナルコンピュータ (以後，PC) 用サイトはスマートフォンで閲覧するには情報量が多い．例えば，そのページで提供している固有の情報の他に，Web サイトにおけるナビゲーションのためのメニューや，広告などが含まれる場合がある．閲覧者が特に情報として欲しいのはそのページ固有の情報であり，スマートフォン用のページでは，それらの情報を判別して，優先的にレイアウトをすることで，見やすいページを構成することができる．しかし，Web サイトによってページレイアウトなどが異なり，ページ固有の情報を自動で判別することは難しい．

本研究では，Web ページ閲覧者向けに，PC 用に作成された Web ページをスマートフォン用に自動変換する方法を提案する．一般に Web サイトは複数の Web ページから構成され，同一サイトの Web ページはレイアウトが類似していることが多い．この点に着目し，閲覧したいページと同一サイトの類似した Web ページとの差分を用いて，スマートフォンでの閲覧において必要な情報を抽出する．

技術的課題として，閲覧したいページと同一サイトの類似した Web ページを取得すること，および，ページ固有の情報と Web ページを構成するのに必要なコンテンツを抽出することが挙げられる．これを解決する方法として，閲覧したいページからリンクされている URL 群を取得し閲覧したいページの URL と比較し，類似ページの候補を得る．さらに，絞られた URL 群のソースを取得し閲覧したいページのソースと比較することで類似した Web ページを取得する．閲覧したいページと類似した Web ページのソースを LCS(Longest Common Subsequence) を用いることで差分抽出し，必要な情報を抽出する．本研究で用いる Web ページは，同一のレイアウト構造の Web ページで構成されているサイトの一部であり JavaScript などで構成されていないことを前提とする．抽出されたコンテンツは，CSS(Cascading Style Sheets) と jQuery Mobile を用いてスマートフォン用にレイアウトを再構成する．自動変換には，プロキシサーバを利用することを想定している．

2　関連研究

Web コンテンツに含まれる無駄な領域を排除する研究に関しては，中村らの漸次的ウェブ閲覧のためのコンテンツ変換が挙げられる [3]．この研究では Web コンテンツに含まれる広告やメニューなどの，目的とするコンテンツとは直接関係の無い領域を排除することで閲覧における無駄を軽減している．しかし，あらかじめ用意されたパターンと一致しないものは排除できない．

携帯端末向けに Web コンテンツを自動変換する研究として，吉川らの携帯端末での閲覧に向けた Web コンテンツ自動変換が挙げられる [4]．この研究では，画像を縮小したり table タグに対して処理を施すことでレイアウトを改善し見やすくしている．

吉川らの研究では画像と対応するタグに対して処理をすることで携帯端末に対応している．しかしながら，コンテンツの量が増えることで閲覧しにくくなる．本研究では，スマートフォンで見やすくするために，それぞれの要素に対して処理を行うのではなく，ページの上部に重要なコンテンツを抽出することで見やすさを実現する．

3　コンテンツ抽出と再構成

本研究で提案する変換方法は次の順で処理を行う．

1. 類似度判定による類似 HTML 取得
2. 差分を利用したコンテンツ解析
3. スマートフォンレイアウト再構成

3.1　コンテンツ整理

スマートフォンの限られた表示領域で必要最低限の情報を表示する必要があるので，コンテンツを 3 つに分類した．ページ固有の情報とサイトを移動するナビゲーションリンクを表示し，共通して存在する情報は閲覧者が任意で表示することで表示する情報を少なくする．

3.1.1　メインコンテンツ

変換対象ページにのみ存在する固有の情報をメインコンテンツとする．利用者はページ固有の情報を得るために閲覧しているので，変換対象ページのコンテンツの中

で一番重要であると考えた．レイアウト構造が等しいソース同士を比較すると，変換ページにのみ存在する固有の情報を取り出すことができる．図 1 では (1) の部分がメインコンテンツである．

3.1.2　共通コンテンツ

変換対象ページと類似ページに共通して存在する情報を共通コンテンツとする．変換対象ページからメインコンテンツを抽出して残ったものが共通コンテンツとなる．会社情報・著作権の記述などの情報が存在する．利用者が閲覧する情報は変換対象ページに固有な情報であり，共通コンテンツは閲覧する頻度が少ないと考えた．図 1 では (3) の部分が共通コンテンツである．

3.1.3　ナビゲーションリンク

変換対象ページと類似ページに，共通に存在するリンクのまとまりをナビゲーションリンクとする．閲覧者が必要に応じて別のページに移動したいとき，ナビゲーションリンクが一番使われる可能性が高いと考えた．ナビゲーションリンクは，共通コンテンツの中に存在するリンクと重複するが利用頻度を考えて別に抽出する．閲覧者の利用頻度が高いリンクをまとめることで閲覧をしやすくする．図 1 では (2) の部分がナビゲーションリンクである．

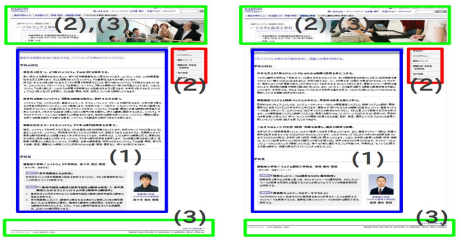

図 1　南山大学ソフトウェア工学科，システム創成工学科のページ

3.2　類似度判定

リンク群の取得

変換対象ページに含まれるリンクを全て抽出する．このとき，相対パスで記述されているリンクを絶対パスに変換する．取得したリンクの URL 中で，最後が .jpg や .gif などで終わっているもの (画像ファイル) などや .pdf ファイル，URL にアンカーが付加されているもの (同ページの特定の箇所に移動するもの) を除く．

類似 URL 候補群の判定

URL 判定では，取得されたリンクの中で変換対象ページと同じサイト，つまり，URL のホスト名が同じものを抽出する．その後，パス名をスラッシュで区切ったものを先頭から順に文字列として比較していき，文字列として一致しなかったものを候補から除外する．最後まで残ったものを URL 判定による類似 HTML 候補群とする．比較の途中で一致する URL がなくなった場合，そこまでで残ったものを類似 HTML 候補とする．

HTML の類似度判定

$$HTML\text{での類似度} = \frac{\text{対象ページと類似ページの共通行数}}{\text{対象ページの総行数}}$$

類似 HTML 候補群に対して LCS を使って変換対象ページと類似ページのソース HTML が何行同じであるかパーセンテージを計算する．類似度が一番高い類似 HTML を類似ページとする．

類似度を求める方法として，ソース HTML，タグのみ，テキストのみ，CSS などのスタイル部分のみの 4 つを検討した．検証した結果，4 つの方法では有為な差がみられなかったことから，特別な処理を施さないソース HTML を採用した．

3.3　差分によるコンテンツ抽出方法

3.3.1　メインコンテンツ

3.2 節で求めた類似ページと変換対象ページとの差分抽出を行なうことによってメインコンテンツを判定，取得する方法を示す．

差分抽出に関して次のような方法がある．

1. ソース HTML の情報をタグのみにしたものから差分を抽出する方法
2. ソース HTML を行単位で比較し差分を抽出する方法

メインコンテンツはそのページ固有の情報を保持している部分であり，テキストが変更されている場合がほとんどである．よって，2 の差分を用いることで効率よくメインコンテンツを抽出してくることができる．

メインコンテンツ抽出の精度を上げる方法として，

1. ソース HTML から body タグ内を抽出
2. 抽出したものからさらに script などを削除

を行う．

この操作を行なうのは，ソース HTML 全文に対して差分抽出を行なった際に head タグ内に含まれている記述は同じであるが，ページを構成する上で重要な部分 (外部 CSS など) を抽出してくることができないからである．これらの部分を取り除いた状態でページを再構成しようとするとレイアウトが崩れてしまう．差分処理に用いない body タグ外の情報と script などに関しては，4 章で示しているレイアウトの再構成で用いる．

整形したソース HTML に対して，行単位での差分を用いることでメインコンテンツを抽出する．同一サイトの Web ページの場合，ナビゲーションリンクなどの記述が同じでテキストのみが異なっていることから差分抽出でメインコンテンツのみ抽出することができる．

3.3.2　ナビゲーションリンク抽出法

変換対象ページからナビゲーションリンクを抽出する方法を示す．複数の Web ページを持つサイトでは，メニューなどを a タグのまとまりで表記されていることが多い．この部分に着目してリンクの抽出を行う．

メニューの例

```
<div id="Topuck-path">
<ul>
<li class="home">
<a href="http://www.nanzan-u.ac.jp/index.html">南山大学ホーム</a></li>
<li><a href="../../Menu/index.html">日本語トップ</a></li>
<li><a href="../index.html">学部・学科</a></li>
<li><a href="../foi.html">情報理工学部</a></li>
<!-- InstanceBeginEditable name="Topic-Path-sub" -->
<li>ソフトウェア工学科：学科の紹介 </li>
<!-- InstanceEndEditable -->
</ul>
</div>
```

複数の a 要素を囲んでいるブロック要素を抽出する．リンクのみを比較したいので，ブロック要素の中にある a 要素以外を取り除く．抽出したブロック要素群をリンク候補群とする．変換対象ページと類似ページのリンク候補群を比較し，一致したブロック要素をナビゲーションリンクとし抽出する．

4 レイアウトの再構成

抽出した各要素を図 2 のレイアウトにしたがって再構成する．利用者は，ページ固有の情報を閲覧するために利用しているので上部にメインコンテンツ配置する．スマートフォンは表示領域が限られていることから，共通コンテンツは通常，非表示にしておき，利用者が表示ボタンを押すことで表示させる．ナビゲーションリンクは外部サイトに飛ぶために利用者が利用するがメインコンテンツや共通コンテンツと比べて重要度は低いと考え最後に配置した．

仮ページにはあらかじめ次のような装飾を行なう CSS と jQuery Mobile を組み込んでおく．

- ページの横幅を端末に合わせる
- 画像を縮小する
- 文字サイズを閲覧し易い大きさに変更する
- 共通コンテンツは一まとめにして表示する

これらの CSS と jQuery Mobile を適用させることによって利用者が閲覧しやすい Web ページを作成する．

図 3 は，南山大学ソフトウェア工学科のページを再構成した結果である，図 1 のメインコンテンツ部分である (1) が上部に配置され，ナビゲーションリンクである (2) は下部にまとめられている．

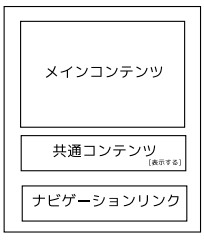

図 2 レイアウトの再構成のテンプレート

5 評価と考察

類似度判定，コンテンツ抽出，レイアウトの再構成，全体の処理に対する検証を行う．

図 3 レイアウトの再構成例

5.1 評価対象の Web ページ

本研究では，検索サイトで単語（スマートフォン，ソフト，大学，レストラン，趣味）を検索し，出てきた重複しない上位 20 件を抽出する．合計 100 件のサンプルサイトを評価対象の Web ページとした．100 件のサンプルサイトは，企業サイト，学校，個人ブログ，Wikipedia，ニュースなどが含まれていた．

5.2 検証方法

類似度判定の検証方法

類似度判定の検証方法は，変換対象ページと類似ページのレイアウト構造が類似しているかを目視で確認する．

差分抽出の検証方法

差分抽出の検証方法は，変換対象ページで判断したメインコンテンツが抽出できているか目視で確認する．

レイアウトの再構成の検証方法

レイアウトを再構成したもので，レイアウトが極端に見づらくなっているものの共通点を上げる．

全体の処理の検証方法

全体の処理の評価方法は，

- 類似度判定のときに取得した URL 数
- 類似度判定の HTML ファイルのサイズ
- time コマンドで real・user・sys の時間

の値を求め処理時間の検証をする．

5.3 検証結果

類似度判定の検証結果

変換対象ページと類似度判定で類似ページだと判断されたもので同一のレイアウト数は 100 件中 68 件である．

差分抽出の検証結果

メインコンテンツは変換対象ページと類似ページの差分であり，変換対象ページで判断したメインコンテンツが抽出できたのは 72 件である．

レイアウトの再構成の検証結果

極端にレイアウトがくずれていたものは 100 件中 10 件であった．レイアウトがくずれているものの特徴として，改行がされていないものや script が表示されてしまうこ

とがあった．原因として，差分抽出をおこなうときにタグや script が動作する部分を抽出しきれないことからくずれてしまうことが分かった．

全体の処理の検証結果

類似度判定のとき，取得する HTML ファイル数は最大 10 個，平均 9 個である (標準偏差 2.23)．

表 1　類似度判定の処理時間の検証結果

	平均	標準偏差	中央値	最大値	最小値
real(秒)	6.3	4.8	4.6	22.8	1.2
user(秒)	1.3	1.6	0.8	12.4	0.4
sys(秒)	0.2	0.3	0.5	1.7	0.2

表 2　全体の処理時間の検証結果

	平均	標準偏差	中央値	最大値	最小値
real(秒)	7.2	5.4	5.4	28.1	1.4
user(秒)	1.9	2.6	1.2	21.8	0.4
sys(秒)	0.7	0.3	0.6	1.7	0.2

表 2 と表 1 から次のことが分かった．

- 表 1 の user と sys の合計がソース HTML を解析して類似 URL 候補群の判定と HTML の類似度判定に要した時間であり，1.5 秒である．
- 表 1 の real から user と sys の合計を引いた時間が通信時間と考えられ，5.3 秒である
- 表 2 の user と sys の合計から表 1 の user と sys の合計から引いた時間が差分抽出とレイアウトに要する時間であり，1.0 秒である．

通信時間が多くかかっているので，これを改善することで処理時間を短縮できる．

5.4　考察

5.2 節で示したように，提案した類似度判定により，100 件の Web ページに対して 68 件の Web ページが同一レイアウトを持つ類似ページを取得できた．68 件のうち 51 件 (78%) が差分抽出によりメインコンテンツを抽出できた．同一レイアウトが持つ類似ページが取得できなかった 32 件のうち 22 件 (69%) が差分抽出によりメインコンテンツを抽出できた．同一レイアウトを持つ類似ページを取得できなかった理由は，変換対象ページと類似ページで広告・メニューの配置が異なっていたり，サイトのトップページだけレイアウト構造が異なることがあげられる．レイアウトの異なるトップページはコンテンツの要素が少ないことが多く，同一レイアウトを持つ類似ページが取得できなくても，CSS や jQueryMobile を使用することで見やすくすることができた．100 件中 27 件は差分抽出できなかった．その理由として，類似ページの取得ができなかった (12 件)，変換対象ページと類似ページがほとんど同じであった (6 件)，変換対象ページと類似ページで言語が異なっていた (5 件)，サイト全体が script で構成されていた (2 件)，サイト全体が table で構成されていた (1 件)，サイト全体が form で構成されていたこと (1 件) があげられる．

本研究で前提とした，同一レイアウト構造をしている Web ページはサイトのトップページからリンクを数回辿ったページに多く，これらのページは 75%は変換できた．サイトのトップページなどは他のページと類似していないこともあるが，69%は変換可能であることを確認した．

今回の検証では，類似度判定の候補は最大で 10 個取得することにしたが，取得数を減らすことで通信時間を削減できる．取得数を最大 3 個にして検証したところ，real の平均が 4.2 秒，user の平均が 1.6 秒，sys の平均が 0.5 秒であり，通信時間を 2.1 秒に短縮できた．3 個にすると，10 個では変換できていた Web ページのうち 4 件が変換できなくなった．最大 5 個にした場合は，10 個の場合と変換結果は同じであった．今後の課題として，精度を落とさず時間を短縮することができるソース HTML 取得数を決める必要がある．

本研究の手法をもちいて自動変換した Web ページとスマートフォン専用サイトを比較したところ，共通コンテンツとナビゲーションリンクの配置・装飾が専用サイトより劣っていた．具体的に，スマートフォン専用サイトはリンクをロゴなどを用いてコンパクトにまとめていたが，本研究ではナビゲーションリンクをコンパクトにまとめることができなかった．これらを改善する方法として，抽出するコンテンツの分類を細分化し抽出することで細かく設定することができる．

6　おわりに

本研究では，スマートフォンでの閲覧に適した Web ページの変換に関する研究を目的としてコンテンツの抽出とレイアウトの再構成に取り組んだ．とくに，差分を用いたコンテンツ抽出を行うことで変換対象ページの固有の情報を抽出することを提案した．類似したページを使って，既存の研究とは違う視点からのコンテンツ抽出に関して実装を行なった．スマートフォン専用サイトはスマートフォンでの見やすさを考慮してあるので，本研究の自動変換した Web ページが専用サイトにより類似することで変換精度が高まったと言える．今後の課題として，処理時間，コンテンツ抽出の細分化，プロキシサーバでの実現があげられる．

参考文献

[1] ASAHI INTERACTIVE, Inc., “スマホユーザーの 9 割が PC サイトの閲覧に不満–急速に浸透する LINE, ”http://japan.cnet.com/news/business/35023271/, 2012.

[2] metaphase.Co.,Ltd., “≪メタフェイズ独自調査≫【第 2 回】企業 Web サイトにおけるスマートフォンサイトの対応状況に関する調査, ”http://www.metaphase.co.jp/press/2011/1212.php, 2011.

[3] 中村聡史, 水口充, 田中克己,“漸次的ウェブ閲覧のためのコンテンツ変換,”情報処理学会研究報告. データベース・システム研究会報告 2005(6), 71-78, 2005-01-20

[4] 吉川裕章, 内田理, 中西祥八郎, “携帯端末での閲覧に向けた Web コンテンツ自動変換, ”情報処理学会研究報告. マルチメディア通信と分散処理研究会報告 2004(22),217-222, 2004-03-04