2I1-4

Webページにおけるレイアウト情報を考慮したDOMの拡張

Extending a DOM Tree based on a Web Page Layout

浅見昌平[*1] 伊藤太樹[*1] 大囿忠親[*1] 新谷虎松[*1]
Shohei ASAMI Taiki ITO Tadachika OZONO Toramatsu SHINTANI

[*1]名古屋工業大学大学院工学研究科情報工学専攻
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology

The Document Object Model (DOM) is API for operating HTML Elements. DOM can't be directly used to represent semantics on web pages. In this paper we propose a method for extending a DOM tree by using semantic information of a layout of a web page. The method can classify and analyze web contents that serves for certain functions. The contents are classified into three categories. We have some rules for automatic classification. The experiments results show that the method can be used effectively.

1. はじめに

近年，Web をユーザがカスタマイズ可能なパーソナライズド技術が注目を浴びている．iGoogle[*1]に代表されるパーソナライズドページでは，ユーザが Web パーツを組み合わせて自分好みの Web ページを構成することができる．特に，ニュースサイトの RSS リーダ，音楽サイトのランキングなど，他のサイトの情報を確認できる Web パーツが人気を集めている．Han らは，他の Web ページ中から欲しい情報のみを抽出し，Web パーツとしてパーソナライズドページに組み込むことを可能にした [Han 07].

Web ページからユーザが求める情報を抽出するためには，HTML の意味的構造を理解しなければならない．なぜなら，人が視覚的に認識するメニュー，ニュース記事，ランキングなどの意味的な情報は，通常 HTML 文書には記述されないからである．これまでにも HTML から意味的な構造を抽出するための研究が盛んに行われている．Chen らは，Web ページを機能別にいくつかのオブジェクトに分類するモデルを提案した [Chen 01]．また，私たちの研究では人が認識する Web ページのコンテンツを抽出するアルゴリズムを提案し，携帯電話への適合を目指した [伊藤 08].

本研究では，Web ページ中のコンテンツに対して意味的な分類を行う．各コンテンツが持つハイパーリンクの数やレイアウト情報を基に，コンテンツが果たす役割を意味付ける．1 つ 1 つのコンテンツが果たす役割を理解することで，ユーザが求める情報を探しやすくなる．

2. Web ページのコンテンツ分類

Document Object Model (DOM) は，Web ページを構成する HTML 要素を操作するための API である．DOM にアクセスすることで描画した際のレイアウト情報を取得することができるため，通常，構造情報しか記述されない HTML に対して付加的な情報を読み取れる可能性がある．

本研究では，文献 [Chen 01] に基づき，Web ページ中のコンテンツのカテゴリを次のように定義する．

- Navigation: 他のコンテンツ，または他のページへのハイパーリンクを持つコンテンツ．例えば，メニューバー，ニュース記事のヘッダなど．
- Information: 意味的な内容を提供するコンテンツ．例えば，ニュース記事，写真など．
- Interaction: ユーザとのインタラクションを提供するコンテンツ．例えば，フォーム，ボタンなど．

連絡先: 浅見昌平，名古屋工業大学大学院工学研究科情報工学専攻，〒 466-8555，愛知県名古屋市昭和区御器所町，asami@toralab.ics.nitech.ac.jp

*1 http://www.google.co.jp/ig

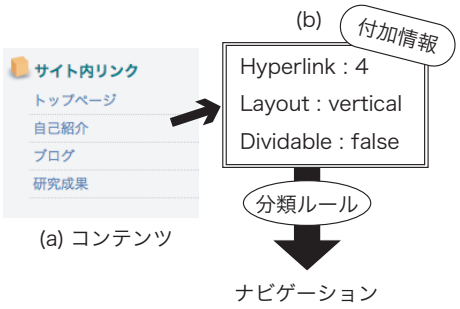

図 1: コンテンツの分類例：ナビゲーション

ここでは，DOM 木におけるブロック要素について，コンテンツのカテゴリ分類を適用する．文献 [Chen 01] 中ではインライン要素についても装飾，表示方法などの属性を付加しているが，その情報は DOM 木から得られるため本研究では扱わない．

図 1 はコンテンツを Navigation に分類する例である．図中の (a) は，Web サイト内の Web ページへ導くための，DIV タグで構成されたブロック要素である．(a) は垂直に並んだハイパーリンクを 4 つ持ち，その内部にブロック要素を持たない．このような付加情報 (b) は，DOM 木を階層的に解析することによって得られる．

本手法では，前処理として DOM 木から得られるレイアウト情報，階層構造の情報を解析し，DOM 木の各ブロック要素に情報を付加する．そして，付加情報を基に分類ルールを用い，Web ページ中のコンテンツに対してカテゴリの分類を行う．

3. DOM 木の拡張と分類ルール

3.1 ブロック要素の属性拡張

コンテンツを分類するために，ブロック要素が内包する要素の情報を DOM 木解析によって求める．構造的な情報は，ブ

表 1: 実験に使用した閾値

hasHyperlinks	2
hasInteraction	1
hasHypertext / hasWord	0.6
ImagePart	0.7

ロック要素が内部に持つハイパーリンクの数 (hasHyperlinks)，フォーム部品の数 (hasInteraction)，ハイパーリンクの文字数 (hasHypertext)，全ての文字数 (hasWord) である．構造的な情報は，DOM 木を階層的に辿ることで取得できる．

ブロック要素のレイアウトから得られる情報は，ブロック要素が内部に持つハイパーリンクの並び (Layout)，画像の表示割合 (ImagePart)，ブロック要素が複数のブロックに分割可能か (Dividable) である．ハイパーリンクの並びは，x 座標が同一である場合 vertical，y 座標が同一である場合 horizontal，1 つでもそろっていない場合 undefined となる．画像の表示割合は，ブロック要素が持つ画像の表示面積の合計を，ブロック要素の表示面積で割った値である．

上記の値を DOM 木の各ブロック要素に属性として拡張する．拡張した属性は，コンテンツを Navigation，Information，Interaction へ分類する手がかりにする．分類対象は，Dividable 属性が false のブロック要素，および内部に複数の種類のコンテンツが複合していないブロック要素である．

3.2 分類ルール

コンテンツをいくつかのルールを用いて分類する．分類ルールは，各コンテンツの特徴を考慮し，閾値によってパラメータ調整ができるように作成する．

Navigation カテゴリに属する要素は，一般的に並列なハイパーリンクを複数持ち，画像やハイパーリンク以外の文字が少ないブロック要素である．よって，hasHyperlinks 値が閾値以上あり，Layout 属性が vertical または horizontal，hasHypertext/hasWord 値が 1 に近いブロック要素を Navigation とする．ハイパーリンクが 1 つしかない場合でも Navigation と判断するかどうかは閾値によって調整する．

Information カテゴリに属する要素は，画像，ハイパーリンク，装飾された文字など，様々な HTML 要素を持つブロック要素である．Navigation との違いは，他の Web ページへと導く役割以外にもユーザへ意味的な情報を提供する点である．よって，hasHypertext/hasWord 値が小さいこと，もしくは ImagePart が大きいことが条件である．

Interaction カテゴリに属する要素は，フォーム部品である INPUT 要素，および SELECT 要素を持つブロック要素である．よって，hasInteraction が閾値以上ある場合に Interaction へと分類する．

これらの分類ルールを用いて，Web ページ中のブロック要素を分類することができる．また，分類されたブロック要素は，Web ページにおける役割を持つ最小のコンテンツである．

4. 実験

図 2 は YAHOO!JAPAN のトップページ*2を対象に分類を行った結果である．表 1 に実験で使用した閾値を示す．

図中の (a) のように，赤枠で囲まれた部分は Navigation である．(a) は，複数のハイパーリンクを持ち，他の Web ページへと導く目的で使われている．また，他の Navigation を検

*2 http://www.yahoo.co.jp/

図 2: コンテンツの分類結果

証すると，内部に画像や，ハイパーリンク以外の文字が含まれている場合でも分類が成功している．図中の (b) のように，緑枠で囲まれた部分は Information である．分類結果を見ると，主にハイパーリンク以外の文字，画像，Flash で構成されたブロック要素が Infomation に分類されている．図中の (c) のように，青枠で囲まれた部分は Interaction である．ユーザが Web ページに対して入力，操作を行う部品を含むブロック要素が検出されている．枠がついていない部分については，ブロック要素以外の HTML 要素で構成されているため，分類が行われていない．

5. まとめ

本研究では，Web ページ中のコンテンツが果たす役割を分類した．分類するカテゴリは，他の Web ページを導くための Navigation，意味的な内容をユーザに提供する Information，およびユーザが入力，操作するための Interaction である．DOM 木から得られるレイアウト情報や構造上の情報を基に DOM 木の拡張を行い，その上でルールを用いて分類を行った．実験では，閾値を調整することでコンテンツの適切な分類結果が得られた．

参考文献

[伊藤 08] 伊藤太樹，近藤圭佑，浅見昌平，大囿忠親，新谷虎松：“携帯電話における Web コンテンツ閲覧のためのコンテンツ抽出アルゴリズムについて，” 第 70 回情報処理学会全国大会講演論文集 (CD-ROM)，2008.

[Chen 01] Chen, J., Zhou, B. Shi, J., Zhang, H., and Fengwu, Q.: “Function-Based Object Model Towards Website Adaptation,” In Proceedings of the 10th International Conference on World Wide Web, pp. 587-596, 2001.

[Han 07] Han, J., Han, D., Lin, C., Zeng, H., Chen, Z., Yu, Y.:“Homepage live: automatic block tracing for web personalization,” In Proceedings of the 16th International Conference on World Wide Web, pp. 1-10, 2007.