5ZD-07

HTML5のセクショニング要素自動付与システムの提案

今泉 智博 † 斎藤 裕佑 †† 西山 裕之 †

東京理科大学理工学部経営工学科 † 東京理科大学大学院理工学研究科 ††

1 はじめに

1990年代後半以降, 情報の発信及び収集のための手段としてインターネットが普及している. 総務省によれば, 2011年末の国内インターネット利用者人口は9,610万人で, 普及率は79.1%である [1]. また別の調査では一日に30分以上インターネットを利用する人が全体の85%に達しているとの結果がある [2]. またインターネットを構成する主要な要素に World Wide Web(以下 Web) がある. 1993年に世界全体で130件しか存在しなかったウェブサイト数は, 2012年には約7億件にまで増加した.

そこで近年 Web ページから主要コンテンツを抽出するニーズが高まっている. ここでの主要コンテンツは, ニュースサイトやブログの投稿ページにおける本文部分のことを指す. その理由として, 既に述べたように日々大量の情報が Web 上に蓄積されていく一方で, 1人のユーザーが到達できる情報量には限界があるという背景が挙げられる. 例えばニュースやブログの本文を抽出することで, 内容を要約し短時間で把握が可能となる. また検索エンジンを利用する際, 広告やナビゲーションなどの不要なコンテンツを検索結果の中に反映することがなくなり, 検索結果の質の向上が期待される. さらに通信時の負担を削減するために, 主要なコンテンツのみを優先して取得することで, Web ページの表示待ち時間を削減することが可能である.

主要コンテンツ抽出手法についての既存の研究では, 吉田ら [3] が同一 Web サイトにおいて他の Web ページには共通して現れない部分を本文箇所として推定する手法を提案している. また Pappas ら [4] は多様な Web ページで主要コンテンツ部分を推定するため, 独自アルゴリズムによる手法を提案している.

2 HTML5 とセクショニング要素について

Web ページを構成するのは HTML タグと呼ばれる識別子によって装飾された文章である. 本研究では主要コンテンツの推定に有用な情報として, HTML5 のセクショニング要素と呼ばれる HTML タグに注目する. HTML5 とは, Web 利用者にとってのユーザビリティ向上を目的として, 2014年10月28日に行われた HTML の5回目の大幅な仕様変更である. またセクショニング要素は HTML5 にて新しく定義された HTML タグで, Web ページ内の文章に対する意味付けを行うことを目的としている. 具体的には次の4つの HTML タグを指す.

- article タグ: Web ページの中で完全もしくは自己完結した構造を表す. これは, フォーラムの投稿, 雑誌や新聞の記事, ブログのエントリ, ユーザーの投稿コメントなどが考えられる. つまり article タグは, 主要コンテンツそのものを表す HTML タグと言える.
- section タグ: 文書またはアプリケーションの一般的セクション (主題を表すコンテンツのグループ) を表す. 通常1つの article タグが0個以上の section タグを内包している場合が多い.
- nav タグ: 他のページへのリンク, またはナビゲーションリンクをもつ部分を表す.
- aside タグ: 周囲のコンテンツとわずかに関連し, かつそのコンテンツから分離すると見なすことができる部分を表す. 具体的にはリード文またはサイドバーのような部分や広告などを指す.

上記のように, セクショニング要素は主要コンテンツの推定に大いに役立つと考えられる. しかし正式勧告されたばかりの仕様のため, セクショニング要素を持たない Web ページも未だ数多く存在する. それらを手動で更新, 修正するのは作業者にとって負担である.

そこで本研究では, HTML5 への更新作業時の負荷軽減と Web におけるユーザビリティの向上を目的とし, セクショニング要素を持たない Web ページに対してそれらを自動付与するシステムを提案する.

なお以降でセクショニング要素を含む Web ページを HTML5 ページ, それらを持たないページを非 HTML5 ページと表現する. また本研究で利用するセクショニング要素は, 4つのうち article タグのみに限定する.

3 システム設計

本システムの構成は図1の通りである. 具体的な処理の流れは以下の通りである.

1. 予め指定した URL の HTML5 ページと非 HTML5 ページのページデータを取得する.
2. 取得した Web ページは用意したデータベースへテキストデータとして保存する.
3. 各 HTML5 ページについて DOM ツリーを生成する.
4. セマンティック要素のノードと, その兄弟ノードや祖先の兄弟ノードについて特徴量を算出する. 例えば図2のような DOM ツリーの場合, 緑のノードを正事例, 赤のノードを負事例とする.

Proposal for a system to automatically append HTML5 sectioning elements to Web pages

Tomohiro Imaizumi†, Yusuke Saito††, Hiroyuki Nishiyama†

†Faculty of Science and Technology, Tokyo University of Science

††Graduate School of Science and Technology, Tokyo University of Science

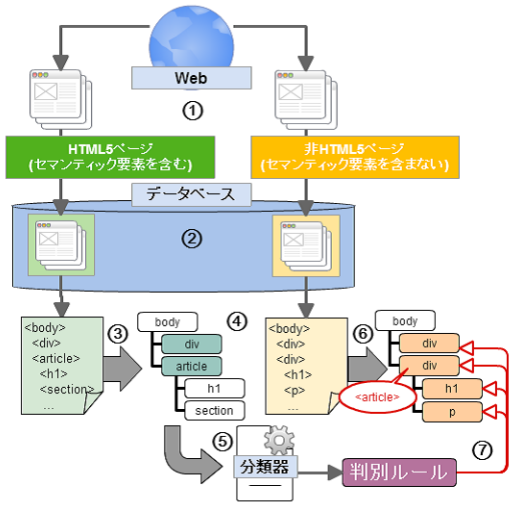

図 1　セクショニング要素自動付与システムの構成図

5. 前述の特徴量から, ノードがセマンティック要素のノードか否かを判別するルールを分類器で生成する.
6. 非 HTML5 ページでも DOM ツリーを生成する. またその後, それぞれの特徴量も求める.
7. 非 HTML5 ページの全ての DOM ノード対して 5 で求めたルールを適用し, 特徴量にもとづいてセマンティック要素のノードとして分類されたノードに対応するタグを article に置き換える.

分類器については SVM やナイーブベイズを用い, 分類器の違いによる性能の比較も行う. また上記の 4 で求める特徴量を求める手順は以下のとおりである. まず次のようなノード N に対するスコア X_{score} を定義する.

- N_{text}: ノード N に含まれる文字数
- $N_{timeinfo}$: ノード N が時間表現を含んでいるか (含んでいる場合 1, 含まない場合 0)
- N_{puncs}: ノード N に含まれる句読点数
- $N_{tag(E)}$: ノード N の子要素に現れるノード E の数

上記のスコアを注目するノード, そのノードが含まれる DOM ツリーの根ノード R についてそれぞれ求める. そして, N_{time} を除く各スコアについて $N_{text}/R_{text}*100$ のように, R のスコアに対する割合を 100 分率にし, これをノード N の特徴量とする. なお N_{time}についてはそのままの値を特徴量として用いる.

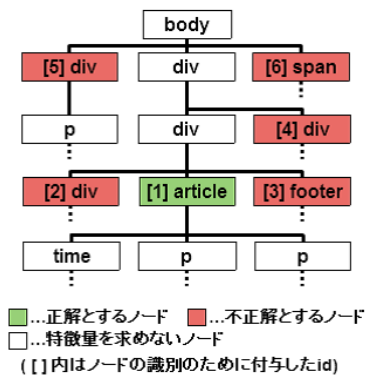

図 2　DOM ツリーで特徴量をカウントするノード

表 1　各ノードのスコア

id	正解	N_{text}	$N_{timeinfo}$	N_{puncs}	$N_{tag(p)}$	$N_{tag(div)}$
1	T	4752	1	148	35	12
2	F	742	0	38	3	0
3	F	15	0	0	0	0
4	F	22	0	0	0	0
5	F	44	0	0	1	0
6	F	0	0	0	0	1

表 2　得られる特徴量の例

id	正解	N_{text}	$N_{timeinfo}$	N_{puncs}	$N_{tag(p)}$	$N_{tag(div)}$
1	T	70.8	1	65.5	89.7	70.6
2	F	11.1	0	16.8	7.7	0.0
3	F	0.2	0	0.0	0.0	0.0
4	F	0.3	0	0.0	0.0	0.0
5	F	0.7	0	0.0	2.6	0.0
6	F	0.0	0	0.0	0.0	5.9

模式的な DOM ツリーを図 2 に示す. 今この DOM ツリーのうちスコアを計算した結果が表 1 であったとする. またこの DOM の根ノードのスコアが $R_{text} = 6710$, $R_{timeinfo} = 1$, $R_{puncs} = 226$, $R_{tag(p)} = 39$, $R_{tag(div)} = 17$ であった場合, これらから表 2 の特徴量を得る.

このようにして得られた特徴量からセクショニング要素を示すノードと, それ以外のノードの分類ルールを生成する. 分類ルールの生成にはサポートベクターマシンやナイーブベイズ分類器などを用いる.

その後生成したルールは非 HTML5 ページから生成した DOM ツリーの全てのノードに対して適用し, 特徴量により各ノードをそれぞれセクショニング要素とそれ以外に分類することで, 最終的にセクショニング要素に相当するノードの推定を行う.

4　おわりに

本研究では, HTML5 への更新作業負荷軽減と Web 利用者にとってのユーザビリティの向上を目的として, HTML5 のセクショニング要素を自動付与するシステムを設計した. 提案したシステムでは, セマンティック要素を持たないページに対し, 適切な位置にセマンティック要素をを付与することに成功した. 本研究では付与できるセクショニングタグが article タグに限定されているが, 今後は他の 3 つのタグについても同様に自動付与できるシステムの作成を目指す.

参考文献

[1] 総務省 平成 24 年版 情報通信白書, 2012
[2] 一般財団法人経済広報センター 情報源に関する意識・実態調査報告書, 2013
[3] 吉田光男, 山本幹雄. (2009). 教師情報を必要としないニュースページ群からのコンテンツ自動抽出. 日本データベース学会論文誌, 8(1), 29-34.
[4] Pappas, N., Katsimpras, G., Stamatatos, E. (2012, September). Extracting informative textual parts from web pages containing user-generated content. In Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies (p. 4). ACM.