Web ページの構造解析とメタデータ候補の抽出

船山 弘孝　　渋田 和宏　　柴田 知秀　　黒橋 禎夫
京都大学大学院情報学研究科
{funayama,shibuta,shibata,kuro}@nlp.kuee.kyoto-u.ac.jp

1　はじめに

Web ページにはサイト名やタイトル、情報発信者のようなメタデータが存在する。Web ページに対する検索や分類を高度化するにはこのようなメタデータが非常に重要である。例えば、国立国会図書館が運営する WARP(Web Archiving Project[1]) では、収集・保存した Web ページに対してメタデータを人手で付与し、検索インタフェースを提供している。しかし、大量の Web ページに対してこれらのメタデータを人手で付与することは非常に高コストであり、自動化が望まれる。そこで本研究では Web ページからサイト名、タイトル、発信者のメタデータを抽出する。

メタデータはページの上部や下部等に出現しやすく、本文には出現しにくいなどの性質がある。そこで提案手法ではページの構造解析を行い、ページ中の領域にフッター領域や本文領域などの領域名を付与する。そして領域情報を利用してサイト名、タイトル、発信者それぞれに対応する処理を行い、各メタデータを抽出する。

2　関連研究

Web ページの構造解析に関する研究として、Cai らは背景色やレンダリング時のタグ間の距離などの視覚的な情報を用いて Web ページを意味のあるブロックに分割する VIPS というアルゴリズムを提案した [1]。Song らは VIPS アルゴリズムで分割したブロックに対してそのブロックの重要度を計算する手法を提案し、人間による評価と同程度の精度を実現している [2]。

メタデータのうちの発信者抽出に関する研究として、加藤らは Web ページの著者とサイト運営者のそれぞれを抽出している [3]。加藤らの手法は、Web ページから発信者がよく表れるような領域に特徴的な条件を満たす文字列を抽出する。次に抽出した文字列から一定のルールに基づく形態素列を発信者候補として抽出し、その中から機械学習を用いて発信者を推定している。また Zheng らは、視覚的な情報を用いて自己紹介ページから著者の抽出を行っている [4]。

3　問題の定義

本稿で扱う問題を次のように定義する。入力となる HTML ファイルおよびそのサイト内のページに対して構造解析を行い、その結果を利用してサイト名、タイトル、発信者の各メタデータを抽出する。ここで各メタデータを以下のように定義する。

サイト名 ホームページの名前やブログ名など各サイト固有の名前。必ず存在する。

タイトル サイト内の個別のページで扱っている内容のタイトル。存在しないページもある。

発信者 Web ページを公開しているサイトの運営者であるサイト運営者とページ内の情報の著者の両者を合わせたもの。ページに明記されておらず不明のページもある。

ここで、サイト名とタイトル、発信者とタイトルは互いに異なり、サイト名と発信者は一致する場合もあるとする。

4　Web ページの構造解析と領域判定

メタデータが出現しやすいヘッダー領域やフッター領域、出現しにくいリンク領域や本文領域などを検出するために Web ページの構造解析を行う。

まず、入力の HTML ファイルを DOMTree に変換する。次に div や table などのページ内の各 HTML タグについて、繰り返し構造を付与する。この情報は領域判定の際や、各メタデータを抽出する際の重要な手掛かりになる。次に Web ページを意味のある単位に分割する。DOMTree を body タグからトップダウンに見て、自分以下のテキスト量がページ全体の閾値[2]以下になるタグを見つけ、そこで分割する (以降、この 50%を切るタグをルートとする部分木をブロック

[1] http://warp.ndl.go.jp/search/

[2] 閾値はページ全体の文字数が 6000 文字未満のページではページ全体の文字数の 50%、6000 文字以上のページでは 3000 文字とした。

表 1: 領域の役割と判定のアルゴリズム

基本ブロック	役割	判定のアルゴリズム
footer	ページ下部に位置する copyright やメニュー等をまとめた領域	ブロック内に「Copyright」「HOME」等特定の文字列を含む ブロック開始がページ末尾から 300 文字以内 ブロック終了がページ末尾から 100 文字以内　(以上全てを満たす)
header	ページ上部に位置するサイト名やメニュー等をまとめた領域	index.*への内部リンクをもつ ブロック開始がページ先頭から 100 文字以内 ブロック終了がページ先頭から 300 文字以内　(以上全てを満たす)
link	関連リンク等、リンクをまとめて記述してある領域	自分以下の全ノードのうち a タグを含む繰り返し構造をもつ、またはその一部である割合が 66%以上
image	画像領域	自分以下の葉ノードのうち 80%以上が img タグ
maintext	ページの主要コンテンツとなる領域	テキストの長さが 200 文字以上 句読点または「の」以外の助詞の全形態素に占める割合が 5%以上 (以上いずれか満たす)
form	検索・コメント送信フォーム等	ブロック以下に form タグを含む
unknown	ページ中の小見出し等	上記以外の領域

サブブロック	役割	判定のアルゴリズム
profile	ページ管理者の名前や生年月日等、プロフィールが記述されている領域	「プロフィール」「ユーザ名」「名前」など特定の文字列を自分以下のブロックに 2 個以上含む
address	電話番号や住所等の連絡先がまとめられている領域	「住所」「連絡先」「ＴＥＬ」など特定の文字列を自分以下のブロックに 2 個以上含む

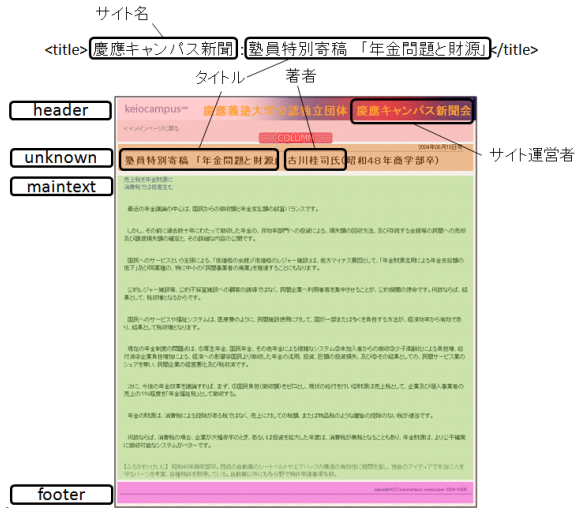

図 1: 構造解析を行った Web ページの例

と呼ぶ)。

最後に各ブロックの領域名を判定する。判定する領域名として表 1 の基本ブロック 7 種類と、サブブロック 2 種類の計 9 種類を考える。このような分類を行うのは基本ブロックがレイアウトを表すような領域であるのに対して、サブブロックは基本ブロック中の文の性質を表すためである。領域の定義とその判定のアルゴリズムを表 1 に示す。基本ブロックに属する領域は表 1 の上から順に条件に適合するかしないかを判断する。その際、header、footer はそれぞれページの上部、下部に位置しなければならないなど、レンダリング時の位置情報も用いている。すべてのブロックに基本ブロックの領域名が割り当てられたのち、サブブロックに属する領域の判断を行う。図 1 に Web ページとその領域判定を行った結果を示す。

5　メタデータ候補の抽出

領域判定結果を用いて、メタデータ候補を抽出する。まず各メタデータが出現しそうな領域 (6、7 節で後述、抽出するメタデータ候補により異なる) からメタデータ候補を抽出する。次に抽出した文字列に対して前処理を行う。まず、文字列を以下の例のように HTML タグやデリミタ (“／” や “:” など) で分割する。

<a> 京都大学 案内
→ 京都大学, 案内

京都大学／ Kyoto University
→ 京都大学, Kyoto University

次に、分割した文字列のフィルタリングを行う。例えば、末尾の形態素が句点・読点などの候補や、“コメント”、“トラックバック”、“名無しさん” などのストップワード[3]をフィルタリングすることで文や Web 特有の高頻度語、一般名詞などを候補から除外する。

6　サイト名・タイトルの抽出

サイト名はサイト内で共通でありタイトルは各ページで異なる場合が多い。また両者の少なくとも一方は title タグに含まれることが多いため、サイト内の複数ページの title タグの共通部分がサイト名、非共通部分がタイトルに対応すると考えることができる。

例えばサイト名が日本抗加齢医学会のサイトにおいて、同一ディレクトリにあるページの title タグの文字列が以下のようになっているとする。

[3]Web 文書 100 万ページに対して構造解析を行い、その候補として抽出された頻度上位 300 件と人手によって整理したものをストップワードに追加した。

表 2: ページ間の title タグの関係とサイト名・タイトルの抽出方法

title タグの関係	サイト名	タイトル	確信度
(1) 兄弟と部分一致	共通文字列	非共通文字列	◎
(2) 兄弟と完全一致または兄弟なし			
(2-1) 親子と部分一致	共通文字列	非共通文字列	◎
(2-2) 親子と完全一致	共通文字列	ページから (*2)	○
(2-3) 親子と不一致	ページから (*1)	title タグ文字列	○
(3) 兄弟と不一致	ページから (*1)	title タグ文字列	○

日本抗加齢医学会 | アンチエイジングとは …(*)
日本抗加齢医学会 | 学術集会・講習会のご案内
日本抗加齢医学会 | 会員の皆様へ
⋮

この例の (*) が解析対象ページの場合、title タグの共通部分である “日本抗加齢医学会” ・非共通部分である “アンチエイジングとは” がそれぞれサイト名・タイトルに対応する。そこでこのような手がかりを利用するために、解析対象ページと、解析対象ページからリンクが張られている同一ディレクトリ内に存在するページ (以降兄弟ページと呼ぶ)、上位・下位ディレクトリに存在するページ (以降それぞれ親ページ・子ページと呼ぶ) の title タグの文字列を比較し、表 2 のようにサイト名とタイトルを決定する。ただし、兄弟または親子ページの少なくとも一方が必ず存在すると仮定している。また表中の “部分一致” とは例のように、前方または後方一致のみを考えている。

6.1 title タグ内の文字列を利用したサイト名・タイトルの抽出

表 2 の (1)、(2-1)、(2-2) のように title タグに共通文字列が存在する場合は、共通部分をサイト名として抽出する。サイト名はサイト内で共通であるので表 2 のように兄弟または親子ページと部分一致する場合は確信度が高いと言える。また (1)、(2-1)、(2-3)、(3) のように title タグが非共通部分をもつ場合はその部分をタイトルとして抽出する。

6.2 ページからのサイト名抽出

表 2 の (*1) の場合は構造解析の結果を利用して解析対象ページから抽出する。header 領域中の文字列、“XX のトップページへ” の “XX” のようなサイト名を含みそうな文字列をサイト名の候補とする。ただしタイトルと同じ文字列はサイト名の候補とは考えない。候補文字列からサイト名を選択する際には “兄弟ページで共通する DOM のパスをもち、同じ文字列” という条件を満たすものを抽出する。

このようにして抽出したサイト名は 6.1 節のような title タグの共通部分を利用して抽出したものよりも確信度が低い。このようなものに対しては、1 節で述べたメタデータ付与を支援するシステムにおいて、人手で修正することが考えられる。

6.3 ページからのタイトル抽出

表 2 の (*2) の場合も 6.2 節と同様に構造解析の結果を用いてページから抽出する。header、unknown 領域中の文字列、h1、h2 タグ内の文字列をタイトルの候補とする。ただしサイト名と同じ文字列はタイトルの候補とは考えない。候補文字列からタイトルを選択する際は、“ページの先頭付近にある unknown ブロックの文字列があれば抽出する” などの複数のルールとその優先度を考え、優先度の高いルールに適合する文字列を選択する。

7 発信者の抽出

7.1 候補文字列を抽出するページ

発信者は解析対象ページに必ずしも含まれているとは限らない。また、解析対象ページに含まれている場合でも、そのサイトのトップページや自己紹介ページの方が容易に発信者を抽出できる場合がある。そこで解析対象ページに加えて以下の条件を満たすページからもメタデータ候補を抽出し、得られた全メタデータ候補の中から発信者を抽出する。ただし、解析対象ページから内部リンク (同一サイト内のリンク) が張られているページのみとする。

トップページらしいページ 解析対象ページの上位ディレクトリにありルートから深さ 2 階層以内にある最上位の index.html

発信者が記述される可能性が高いページ アンカーテキストが “プロフィール”、“会社概要” などの場合のリンク先のページ

7.2 候補とする文字列とフィルタリング

発信者候補として header、footer、unknown、profile、address 領域中の文字列、title タグ内の文字列、meta タグの name 属性が author、copyright の場合の content 属性の値を抽出し、その中から発信者を選択する。ただし、6 節でタイトルと判定された文字列は発信者候補としない。これらの文字列に対して例えば末尾の形態素が動詞の場合を除くなどのフィルタリングを行う。

表 3: メタデータ抽出の実験結果

	サイト名	タイトル	発信者	
	精度	精度	Recall	Precision
baseline	-	-	0.426	0.429
提案手法	0.800	0.750	0.826	0.169

7.3 名詞句の解析・候補の選定

前節までで得られた文字列のうち以下の条件を 1 つ以上満たす文字列を発信者候補として抽出する。

パターン “XX のホームページ”、“XX の日記” のような発信者が含まれそうなパターンから “XX” を発信者候補として抽出する[4]。

固有表現 発信者は固有表現である場合がほとんどである。そこで固有表現解析の結果、人名、地名、組織名を含む表現を抽出する。

カテゴリ 末尾の名詞のカテゴリ[5]が 組織・団体、場所-施設のものを抽出する。

8 実験

実験には加藤ら [3] の評価で用いられている文書セットのうち、解析対象ページに発信者が存在する 1504 文書を用いた。正解データは各 Web ページに対してサイト運営者と著者のみが付与されている。サイト名・タイトルについては 100 文書をランダムに選択し、人手で正解を付与した。実験結果を表 3 に示す。サイト名、タイトル抽出はそれぞれ 0.80、0.75 の精度であった。発信者抽出においてサイト運営者と著者を区別せずに発信者として評価し、以下の式を満たす場合を正解とみなした。

$$\frac{2 \times (\text{共通部分文字列の長さ})}{(\text{正解文字列の長さ}) + (\text{出力文字列の長さ})} > 0.7$$

ベースラインとして領域情報を用いずに copyright、meta タグの name 属性が author、copyright の場合の content 属性の値を全て抽出した場合と比較した。比較結果より copyright や meta タグの情報を用いただけでは発信者を網羅的に抽出することは不可能であり、構造解析の結果を利用して発信者解析を行うことの有効性を示すことができた。

例えば、図 1 のページから発信者、サイト名、タイトルをそれぞれ抽出すると表 4 のようになる。

サイト名抽出に関する誤りとして、表 2 の (3) のパターンの場合にサイト名が抽出できない原因は、ページの構造解析の際に、サイト名を含む文字列が header 領域などの候補文字列を抽出する領域として判断されないためである。そこで、サイト内のページは同じ DOM 構造をもつことなどを利用して構造解析の精度を向上させる必要がある。

[4] このようなパターンは全部で 10 種類ある。
[5] JUMAN のカテゴリ全 18 種を利用

表 4: 図 1 のページに対する各メタデータの抽出結果

メタデータ	候補文字列	判定
発信者	keiocampus newspaper	正解
	古川桂司氏	正解
	慶應キャンパス新聞会	正解
	慶應キャンパス新聞	
	塾員特別寄稿	
サイト名	慶應キャンパス新聞	正解
タイトル	塾員特別寄稿「年金問題と財源」	正解

また、発信者抽出における Precision 低下の原因として “アガリクス”、“シルバーアクセサリー” などの未知語が固有表現解析の際に組織名などと誤って判定され、適切にフィルタリングが行われないことによるものがある。また “松坂大輔”、“紀伊国屋書店” などの発信者以外の人名や組織名なども原因の一つとしてあげられる。

9 まとめと今後の課題

本稿では、Web ページの構造解析を行い、その結果を用いて 発信者、サイト名、タイトルの 3 つのメタデータを抽出する手法を提案した。

今後の課題として、本稿ではサイト運営者と著者を発信者として区別せずに抽出したが、これらを区別して判断することが上げられる。その際の候補文字列として解析対象ページにリンクしているアンカー文字列を利用することも考えられる。

また複数のページで共通する部分 (Template) をあらかじめ獲得しておき、それを構造解析や各メタデータの抽出の際に利用することで、それぞれの精度が向上すると思われる。

参考文献

[1] Deng Cai, Shipeng Yu, Ji-Rong, and Wei-Ying Ma, Vips: a vision-based page segmentation algorithm, *Technical Report MSR-TR*, (Microsoft Research, 2003).

[2] Ruihua Song, Haifeng Liu, Ji-Rong Wen, and Wei-Ying Ma, Learning block importance models for web pages, *Proceedings of the 13th international conference on World Wide Web, WWW 2004*, (ACM, 2004), pp. 203–211.

[3] 加藤義清, 河原大輔, 乾健太郎, 黒橋禎夫, 柴田知秀, Web ページの情報発信者の同定, 人工知能学会論文誌, 第 25 巻, (2010), pp. 90–103.

[4] Shuyi Zheng, Ding Zhou, Jia Li, and C. Lee Giles, Extracting Author Meta-Data from Web using Visual Features, *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, (IEEE Computer Society Washington, DC, USA, 2007), pp. 33–40.