Vol.2009-DBS-149 No.15
2009/11/21

Webページの構造と内容の分析による手法掲載部分の抽出

野中諒志[†1] 湯本高行[†1]
新居学[†1] 高橋豊[†1]

ユーザがWeb上で手法情報を探すとき，汎用的な検索エンジンでは効率的に発見することは難しい．それに対して，クエリ修正によるアプローチが行われているが，この場合では，トピックに依存した形となり，さらに，手法情報の書かれた部分である手法掲載部分を抜き出すことができない．そこで，本研究では，ページを解析することで，トピックに依存せずに手法情報の掲載部分を抽出する方式を提案する．提案方式では，手順表現や経過表現およびページの構造を用いて手法情報の掲載部分を抽出し，取り出した手法掲載部分の始点と終点をHTMLのタグから判断し修正することで，より正確な手法掲載部分の抽出を行う．また提案方式を用いて実験を行った結果，平均として適合率70%，再現率67%で手法掲載ページが発見できることがわかった．また，著者らの先行研究に比べて，適合率と再現率の両面において20%ほどの数値の上昇が見られた．

Extracting How-to Information Block Based on Analysis of Structure and Content of Web Page

Ryouji Nonaka,[†1] Takayuki Yumoto,[†1] Manabu Nii[†1]
and Yutaka Takahashi[†1]

It is difficult to discover how-to information by conventional Web search engines. In order to solve this problem, most of researchers try to filter Web pages that are not how-to information by query modification approach. However, these approaches can't extract how-to information block of Web pages. Moreover they depend on their topic. Hence, we propose a method to extract how-to information block of Web pages without depending on topics by analysis of Web pages. Our algorithm extracts how-to information by using numbering items and phrases expressing procedure of how-to information in Web pages. Then, we change starting point and ending point of how-to information block by checking description of html tags. In our evaluation we found that average of precision of algorithm is 70% and average of recall is 67%. In addition, we also compare performance of our algorithm and our previous algorithm. We found that precision and recall of our algorithm are about 20% higher than precision and recall of our previous algorithm.

1. はじめに

近年，Web上の情報量が爆発的に増え，それに応じて特定の情報を探す際のユーザの負担は大きくなっている．現在は，Googleの PageRank[1] などの汎用的な検索用途のためのアルゴリズムが数多く開発され，それらを用いた検索エンジンが広く使われている．しかし，これらの汎用的な検索エンジンで特定の情報を探す場合，ユーザがクエリを工夫しなければいけない．また，場合によっては，クエリを工夫するだけでは目的のページに辿り着けず，大量の情報の中からユーザが特定の情報を探さなければいけなくなり，ユーザの負担は大きなものとなる．故に専門的な検索エンジンの開発が重要視される．現在多くの研究で，クエリ修正により情報を絞りこむという手段が取られている．しかし，クエリ修正では，トピックに依存しないような，「物の意味や定義が記された情報」，「人や物の評価や評判が記された情報」，物の作り方などの方法や手段が記された「手法情報」のような1つのジャンルの情報を取り出すことは困難になる．

そこで本研究では，さまざまなジャンルの中で手法情報という1つのジャンルに焦点を置く．ここで，手法情報へと焦点を置いた理由としては，情報爆発時代に向けた新しいIT基盤技術の研究プロジェクトの情報検索に対する信頼性に関する調査[2] により，表1のような結果が見られたからである．この表は，1000人規模のアンケートで「Q4. 検索した時に知ろうとしているもので多い物は何ですか」という問いに対して，回答者が上位5位までを答えた結果の総和である．ここからもわかるように，「方法・手段」が3番目に欲しい情報となっている．ここでの，1番目と2番目にユーザが知りたいと思っている情報は，現在の汎用的な検索エンジンでも網羅できるが，3番目にユーザが知りたいと思っている手法情報は，専門検索エンジンを用いなければ，汎用的な検索エンジンでは取得が面倒になると考えられる．以上のことから，手法情報のみを抽出することはユーザにとって有用なものである

†1 兵庫県立大学大学院工学研究科
Graduate School of Engineering, University of Hyogo

と言える．

表 1 検索時に知ろうとしていること

順位	知ろうとしていること	回答数
1	検索キーワードの意味・定義	657
2	もの	614
3	方法・手段	575
4	場所・地名・地域	535
5	ニュース	463
6	評判・評価	459
7	人・組織	380
8	理由・原因	376
9	コンテンツ	247
10	URL	197
11	イベント	185
12	主張・意見	168
13	時期・時間	137
14	その他	7

また，このような情報を探す際のユーザの負担を軽減する意味で，本研究では手法情報の中でも見やすいページを提示することを最終目標としているが，本稿においては手法情報の抽出方法についてのみを提案している．

提案手法では，トピックへの依存性を取り除くためにクエリ修正という方法を取らずにページの DOM ツリーを解析するという方法を取る．具体的には，DOM ツリーにおいて「1、…。2、…。」のような手順表現や「まず…」「次に…」のような経過表現が書かれているテキストノードに注目することで大まかな手法情報の書かれている部分，つまり手法掲載部分を見つけ出し，取り出した手法掲載部分内での内容の切れ目や内容の塊と思われる部分毎に区切りを入れ，正確な手法掲載部分を見つけ出すというものである．

2. 本研究の位置付け

2.1 研究概要

本研究では，ユーザがクエリを入力したときに，Web からクエリに関する手法掲載ページのみを抽出し，見やすい順にリランキングしたものをユーザに提示することを最終目標としている．

大まかな処理の流れを以下に示すと共に，概念図を図 1 に示す．

(1) ユーザがクエリを入力し既存の検索エンジンから検索結果を取得
(2) 手法掲載ページのみ抽出
(3) 見やすい順にリランキングし，リストをユーザに提示

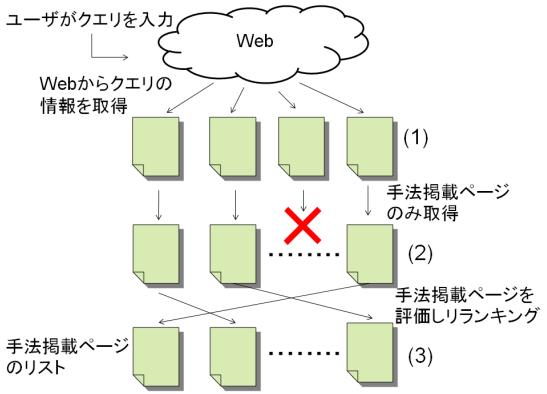

図 1 処理の概念図

ここでの (1) と (2) の部分はクエリ拡張に関する研究で行われていることが多く，(1) の部分は十分にできているが，(2) の部分が十分にできていないことが多い．たとえば，小久保らは，検索隠し味を用いて専門検索エンジンを構築する手法を提案している[5]．例えばユーザが牛肉を使ったレシピについて知りたい場合,「牛肉」と既存の検索エンジンに入力しても，牛肉を使ったレシピが得られることは少ない．しかしここで「塩」というキーワードを追加で入力するとレシピが多く得られる．検索隠し味とは，ここでいう「塩」である．小久保らの手法では，クエリ拡張により専門検索エンジンを構築しているが，これでは OS のインストールや折り紙の折り方といったような多方面での手法掲載ページに対応するとなると，それだけ多様な付属語が必要となり，クエリに対する適切な付属語を決定するために対象ドメイン毎に学習が必要になる．つまりここでもわかるように，クエリ修正からのアプローチではトピックに依存せずに手法情報全般を取得するというのは困難である．また，これらクエリ修正では，手法掲載ページが取り出せたとしても，手法掲載部分を取り出すということはできない．故に，本研究ではページの構造から必要な部分のみを得るという手段

Vol.2009-DBS-149 No.15
2009/11/21

を取る．

また，我々は，すでに見やすい手法情報の取得において (2) と (3) の両方を提案している[3]．しかし，既存の方法 (以下従来手法とする) では手法掲載ページの抽出方法とリランキング方法の両面において問題が見つかった．よって本稿では，(2) を改良した手法について提案する．

従来手法では，ページの分割に VIPS を用い，そこから手法掲載部分を見つけ出し抽出するという方法をとっている．しかし，VIPS を用いた手法掲載部分の抽出方法では，いくつか問題があることがわかった．2.2 の関連研究でも述べるが VIPS には DoC という分割のための尺度が存在する．また，その DoC の最小単位は決まっており，それにより最も細かなレベルの分割でも分割しきれない場合があるということがわかった．また，VIPS は汎用的な分割手法であるがために，手法情報においては分割すべきでない場所で分割する場合があり，正確な手法掲載部分の取得が困難になる．分割しきれない場合の例を以下の図 2 に示す．

図 2 において，VIPS による分割の限界が (a) のような場合に問題が生じる．実際の手法掲載部分である (b) と見比べてもわかるように，(a) では手法とは無関係の部分が含まれている．このような問題が生じると，リランキングの際に正確な判定ができないという問題も発生する．また，このような問題以外にも VIPS では分割が全くできないページが存在することもわかった．故に本稿では，このような問題を解決するために，VIPS を用いない形での手法掲載部分の抽出方法を提案する．

2.2 関 連 研 究

手法掲載部分の抽出に関連して，武智らは HTML 文書の手順に関する箇条書きを抽出する手法を提案している[4]．このシステムでは，HTML 構造の ⟨OL⟩ タグまたは ⟨UL⟩ タグで囲まれた部分を抽出するという方法を用いている．しかし，この表現を用いた場合，⟨OL⟩ タグまたは ⟨UL⟩ タグ以外の手順表現，手法掲載部分を抽出することはできない．そこで本研究では，正規表現により「1、…。2、…。」といった手順表現の他に「…たら…」「次に…」といった経過を表す語にも注目し，手法掲載部分の抽出を行う．これにより，手順表現を用いられていなくても，タグに依存せずに手法掲載部分を抽出できる．

Deng らは，Web ページの構造やページ内の意味のつながり等から Web ページをブロック化し階層化する VIPS というアルゴリズムを提案している[6]．このアルゴリズムでは各ブロックに対して DoC(Degree of Coherence) という内容がいかに一貫性があるかという尺度が用意されている．その DoC の値が大きくなるにつれて内容が密集したものを表し，それに応じて葉ノードが多くなり，階層も深くなる．また，ユーザは閾値 PDoC(Permitted Degree of Coherence) を設定することができ，ページ内のすべての PDoC ≧ DoC となるノードに対して PDoC < DoC になるまで分割を行う．このアルゴリズムでは視覚的な要素等も考慮に入ってくるため，正確に手法掲載部分のみを取得するということは難しく，また VIPS アルゴリズムではページによって分割できないページがある．そこで本研究では，手法掲載部分を正確に取得するために，DOM ツリーと特定のタグ情報からページを分割する．

(a) VIPS による分割　　(b) 実際の手法掲載部分

図 2　分割しきれない例

砂山らは，Web ページ要約のための HTML テキスト分割を提案している[7]．このアルゴリズムは，ブラウザで表示の際に改行を伴い，段落を構成できる内容の塊や句切れを示すタグであるブロックレベル要素毎，リンクの集合毎，文の集合毎に分割を行う．また，このアルゴリズムでは，文であるか否かの判定に「句点の存在」以外に，自立語と付属語や助詞，助動詞の割合も考慮に入れている．このようなアルゴリズムでは，形態素解析を行わないといけないため，高速な処理には向かないと考え，本研究では，文であるか否かの判定に「句点の存在」以外では，格助詞である「を」が存在するかを判定基準に用いている．これ

により形態素解析を行わず，比較的高速な処理が可能になる．また，ブロックレベル要素の一部を本研究ではページに区切りを入れるための要素として用いている．

吉田らは，事前に教師情報を準備する必要のない単純なアルゴリズムで Web ページ群からコンテンツを抽出する手法を提案している[8]．その中でコンテンツを抽出するために，DOM ツリーを判定基準として用いたページの分割方法を提案している．このアルゴリズムでは，Web ページ内の見出しや段落など文書の基本構造を構成するためのブロックレベル要素である ⟨h⟩ タグや ⟨p⟩ タグ，⟨div⟩ タグなどに注目し，これらのブロックレベルの要素を基準としてページを分割している．この方法を用いれば，容易に Web ページを分割することができる．しかし，このアルゴリズムでは，内容の切れ目等の考慮を行っていないため，細かなレベルでの分割を行う際には，正確な分割を行えない．故に本研究では，手法掲載部分を正確に取得するために，このような構造的な分割を行うと共に，内容の切れ目や内容の塊などに注目した分割も行う．

3. 手法掲載ページの抽出方法

3.1 手法掲載ページの分類

本稿では手法掲載ページの定義を「段階を経て物事を説明しており，最終的な結果を掲載しているページ」とする．このような手法掲載ページを見比べ考察した結果，ここでの段階を経た説明とは，以下の 2 つがあることがわかった．

1 ⟨LI⟩ タグや「1、…。2、…。」のように箇条書きや順序を付けて説明する
2 箇条書き等がない場合は「まず〜」や「次に〜」のように経過を表す言葉で説明されている

以上に示した各説明の形より得られた以下に示すような基準で，手法掲載ページであるか否かを判断する．

(a) 「1、…。2、…。」といった順序に沿った説明があるか
(b) 動作や物事の経過を表す語が存在するか

上記での (b) に当たる動作や物事の経過を表す語とは，表 2 に示すような語である．

また，このような説明が行われているページに共通されて見られた以下のような特徴があった．

特徴 手法掲載ページの手法掲載部分内では，文末に過去形表現「〜た。」が用いられることはない

表 2 動作や物事の経過を表す語

語句が現れる位置	語句
文中	〜たら、〜
	〜後、〜
文頭	始めに〜
	まず〜
	最初に〜
	次に〜
	続いて〜
	最後に〜

以上の特徴より，以下の基準も手法掲載ページであるか否かの判断に用いることができる．

(c) 文末に過去形の表現が存在していないか

3.2 手法掲載部分抽出のアルゴリズム

本稿のアルゴリズムは，ページの DOM ツリーと HTML タグを判断基準として手法掲載部分を抽出するものである．アルゴリズムの概要を以下に示すと共に，処理の概要図を図 3 に示す．

(1) ページの DOM ツリーを取得し，手法について書かれているテキストノードのみを取得
(2) 取得した全テキストノードの共通の親ノードを取得
(3) タグを用いた始点と終点の修正
(4) 取得した手法掲載部分の過去形表現の出現割合の判定

このようにして手法掲載部分の有無を判定し，手法掲載部分のあるページを手法掲載ページ，ないページをそれ以外のページとして判断し，手法掲載ページのみを抽出する．

(1) と (2) の詳細を 3.2.1 に，(3) の詳細を 3.2.2 に，(4) の詳細を 3.2.3 に示す．

3.2.1 手法について書かれているテキストノードの取得とその親ノードの取得

手法について書かれているか否かは，3.2.1 の (a) で示した基準「手順表現への注目」と (b) で示した基準「動作を表す語への注目」の判定を行うことで判断する．

また，一般的に，手法掲載ページの手法掲載部分は文章化されて説明がなされている．故に，この処理は文章化された文字列の含むテキストノードにのみ行う．文章化された文字列であるか否かは，以下のような基準を元に行う．

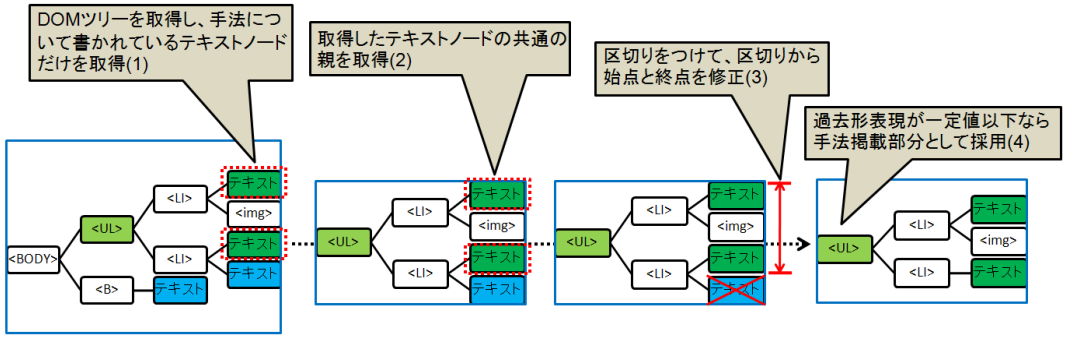

図 3 手法掲載部分抽出の処理の概念図

基準 1 句点が用いられているか
基準 2 文字列内で「を」が使われているか

基準 2 を用いる理由は，主に「を」は格助詞以外では，ほとんど使われず，また手法掲載部分において，文章化されている文字列に対しては，この格助詞「を」が頻繁に用いられるためである．また，このように「を」だけを判断基準に使うことで，形態素解析を行う必要がなく判定できるため，処理を高速化することもできる．また，このようにして得られたテキストノードとそれらの DOM ツリーの関係を用いて，それらの最近傍の親ノードを取得する．

3.2.2 タグを用いた始点と終点の修正

構造的な分割は (1) と (2) で行うので，次は意味的な分割を行い，始点と終点を修正する必要がある．内容が切れる部分と内容の塊ごとで分割を行えば意味的な分割が行える．本研究で用いる，内容の区切りや内容の塊を表すタグを表 3 に示す．このようなタグ毎に区切りを入れ，手法の開始地点の直前の区切りと手法の終了地点の直後の区切りの位置を取得する．その後，その区切りの位置から手法掲載ページの始点と終点を修正する．

3.2.3 取得した手法掲載部分の過去形表現の存在の判定

過去形表現の存在の判定は，3.2.1 の (c) で示した基準「過去形表現の存在」の判断を行い，取得されたブロックに「…た。」という表現がないかにより判断する．3.2.1 でも述べたように一般的に，手法掲載部分の文末において「…た。」のような過去形表現は使われない．

表 3 区切りとして考えられるタグ

タグ名	用いられる箇所
⟨p⟩	文章の区切りとなる場所
⟨ul⟩	内容の一つの塊
⟨ol⟩	
⟨table⟩	

ここでは，このような過去形表現がある文章化された文字列が全体にどれほどあるかという割合が，閾値 θ%以上ある場合，手法掲載部分でないと判断している．

3.3 処理の具体例

手法掲載部分の抽出方法の具体例を図 4 に示す．

図 4 にて，手法について書かれている部分は色の変わっている箇所，つまり B2_1_2_2 と B2_1_1 とする．この場合図 4 の Step 1 より，最近傍の共通の親は B2_1_2 となり，手法とは関係のない B2_1_2_1 も同時に取得してしまうことになる．そこで，このような不必要な部分を排除するために，3.2.2 でも示した表 3 のようなタグに注目しのように区切りを入れる (図 4 の Step 2 参照)．ここで，区切りを入れるための具体例として B2_1 の中身である HTML データの例を図 5 に示す．図 5 では，青文字のテキストが手法と無関係のテキストで，緑のテキストが手法と関連のあるテキストとなっている．また，ここでは，⟨P⟩ タグが区切りを入れるポイントとなるので，ここで区切りを入れる．その ⟨P⟩ タグを区切りの基準として不必要な部分である B2_1_2_1 を取り除いたものを，最終的な手法掲載部分としてのように取得する (図 4 の Step 3 参照)．

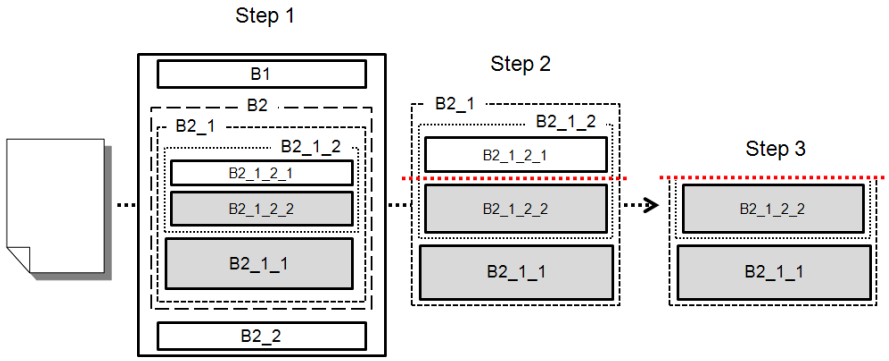

図 4 処理の具体例

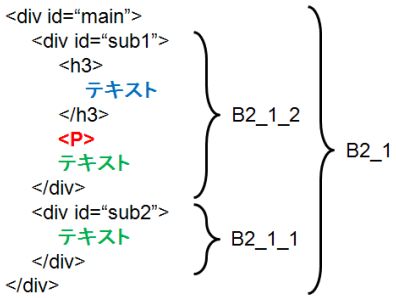

図 5 B2_1 の HTML データ

4. 実　　験

実験は「肉じゃが」,「Fedora Core」,「光る泥団子」の 3 種類のクエリに対して行った．また，評価対象は，各クエリ毎の Yahoo!の検索エンジンの検索結果の上位 50 件である．本実験では過去形表現を判定する閾値 θ は 10%としており，これは，10%以上過去形表現が含まれていれば手法掲載部分でないと判定するということである．これにより求まる，手法掲載ページの数，提案したアルゴリズムにより抽出したページの数，アルゴリズムにより抽出された手法掲載ページの数から各クエリに対するアルゴリズムの適合率と再現率を求める．適合率と再現率を求める具体的な式を以下に示す．

$$precision = \frac{C}{A} * 100 \quad (1)$$

$$recall = \frac{C}{B} * 100 \quad (2)$$

ここでの，*precision* が適合率，*recall* が再現率を表している．また，A がアルゴリズムにより抽出したページの数，B が全体の中での手法掲載ページの数，C がアルゴリズムにより抽出された手法掲載ページの数を表している．適合率と再現率については，正解を以下の 2 通りにした場合について求める．

(a) 手法掲載ページと判断できた場合

(b) 抽出した手法掲載ページで手法掲載部分も正確に抽出できた場合

また，従来手法では Web ページの分割アルゴリズムとして VIPS を用いて，手法掲載部分の発見を行っている．この従来手法を用い，提案手法の実験と同様の実験を行い，提案手法の結果と比較することで，本稿での提案手法の有効性について考察する．

実験結果において，提案手法の (a) の場合の結果を表 4 に，(b) の場合の結果を表 5 に示す．また，従来手法の (a) の場合の結果を表 6 に，(b) の場合の結果を表 7 に示すと共に，提案手法と従来手法との (a) の場合の比較結果を表 8 に，(b) の場合の比較結果を表 9 に示す．

表 4 手法掲載ページ抽出結果の適合率と再現率 (提案手法)

クエリ	適合率	再現率
肉じゃが	0.83	0.77
Fedora Core	0.76	0.73
光る泥団子	0.50	0.50
平均	0.70	0.67

表 5 手法掲載部分抽出結果の適合率と再現率 (提案手法)

クエリ	適合率	再現率
肉じゃが	0.72	0.68
Fedora Core	0.52	0.50
光る泥団子	0.43	0.43
平均	0.56	0.54

表 6 手法掲載ページ抽出結果の適合率と再現率 (従来手法)

クエリ	適合率	再現率
肉じゃが	0.81	0.68
Fedora Core	0.65	0.65
光る泥団子	0.54	0.50
平均	0.67	0.61

表 7 手法掲載部分抽出結果の適合率と再現率 (従来手法)

クエリ	適合率	再現率
肉じゃが	0.31	0.26
Fedora Core	0.38	0.38
光る泥団子	0.31	0.31
平均	0.33	0.32

表 8 手法掲載ページ抽出結果の比較表 (平均値)

	適合率	再現率
提案手法	0.70	0.67
従来手法	0.67	0.61

表 9 手法掲載部分抽出結果の比較表 (平均値)

	適合率	再現率
提案手法	0.56	0.54
従来手法	0.33	0.32

表 8，表 9 を見ても明らかなように，手法掲載ページだと判定できた場合は，従来手法と提案手法とで大きな変化は見られない．しかし，手法掲載部分を正確に取り出す場合には，適合率と再現率の両面において提案手法の方が従来手法より大きく上回っていることがわかる．このような差が出る原因としては，VIPS の汎用的なページ分割にあると考えられる．VIPS では汎用的なページ分割を行うために，手法掲載ページにおいては分割すべき場所でなくても分割する可能性がある．またページによっては，分割が粗いレベルでしか行えない場合もある．このように正確な分割が行えないために，手法掲載部分を取得する場合には，大きな差が出たものと考えられる．

また，提案手法においても以下のような問題点があることが実験からわかった．

まず 1 つ目は，全ての実験結果に見られたことで，手法掲載部分が正確に取得できていないページが何点かあることから，手法掲載部分の抽出方法の 1 つの基準である経過表現が十分に機能していないことが問題であるとわかった．本研究で用いている経過表現は表 2 に示したとおりであるが，これは手法掲載部分で使われる代表的な経過表現であり，網羅的であるとは言い難い．それ故に始点と終点の修正が正確に行えず，表 4 より表 5 の方が全体的に適合率と再現率を下げるという結果になった．しかし，経過表現を多くすると，次は手法掲載部分でない場所で誤判定をするという問題もあり，単純に多くするだけでは，逆に精度が落ちてしまうことは明らかである．故に，この問題を是正するためには，表層的な部分への注目の強化ではなく，パターンマッチングなどの評価も取り入れる必要がある．

次の問題点は，クエリ「光る泥団子」で多くみられたもので，動画を判定することができないという問題である．現在のアルゴリズムでは，テキストに対する判定しか行えず，画像や動画のみで手法が記述されている場合には，手法掲載部分であると判定できない．クエリ「光る泥団子」は，このような動画や画像のみでの説明がなされているページが非常に多く，今回の実験では適合率と再現率を大きく下げるという結果になった．この問題を是正するには，画像解析や動画解析を行うアルゴリズムが必要だと考えられる．

5. おわりに

本稿では，ページの DOM ツリーに対して構造的な解析を行った後に，タグの情報から内容の切れ目を判断することでページ中の手法掲載部分を抽出する方法を提案した．提案手法を用いて実験を行った結果，従来手法より有効な結果が得られることがわかったが，現在定義している経過表現では網羅的でなく，判定することができないページがあることもわかった．経過表現のパターンを単純に増やすだけでは，誤判定を引き起こす可能性が高いため，パターンマッチング等を取り入れた手法掲載部分の抽出方法の開発が今後の課題であることがわかった．

謝辞 本研究の一部は，平成 21 年度科研費基盤研究 (B)(2) 「ユーザの潜在的意図を用いたレス・コンシャス情報検索基盤の構築」(課題番号：20300039) によるものです．ここに記して謝意を表すものとします．

Vol.2009-DBS-149 No.15
2009/11/21

参 考 文 献

1) Sergey Brin, Lawrence Page, “The anatomy of a large-scale hypertextual Web search engine”, Computer Networks and ISDN Systems, Vol.30, Issue1-7, pp.107-117, (1998).
2) “情報検索に対する信憑性に関する調査”, http://www.dl.kuis.kyoto-u.ac.jp/i-explosion/report/index.html
3) 野中 諒志, 湯本 高行, 新居 学, 高橋 豊, “Web ページの表層的な特徴を用いた手法情報の発見”, DEIM フォーラム 2009, C6-3, (2009).
4) 武智 峰樹, 徳永 健伸, 松本 裕治, 田中 穂積, “WWW ページからの手順に関する箇条書きの抽出”, 情報処理学会論文誌, Vol.44, No.SIG 5, pp.1-13, (2003).
5) 小久保 卓, 小山 聡, 山田 晃弘, 北村 泰彦, 石田 亨, “検索隠し味を用いた専門検索エンジンの構築”, 情報処理学会論文誌, Vol.43, No.6, pp.1804-1813, (2002).
6) Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma, “VIPS: A VIsion based Page Segmentation Algorithm”, Microsoft Technical Report, (2003).
7) 砂山 渡, 井山 晃洋, 谷内田 正彦, “重要文抽出による Web ページ要約のための HTML テキスト分割”, 電子情報通信学会論文誌, Vol.J87-D-1, No.12, pp.1089-1097, (2004).
8) 吉田 光男, 山本 幹雄, “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”, DBSJ Journal, Vol.8, No.1, pp.1-6, (2009).
9) “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.net/