2B3-2

Webページ中の部分領域を対象とした検索システム

The search system for the partial area of Web page

田崎 雄一郎[*1] 福原 知宏[*2] 佐藤 哲司[*1]
Yuichiro TASAKI Tomohiro FUKUHARA Tetsuji SATOH

[*1]筑波大学大学院 図書館情報メディア研究科
Graduate School of Library Infomation and Media Studies, University of Tsukuba

[*2]産業技術総合研究所 サービス工学研究センター
Center for Service Recearch, National Institute of Advanced Industrial Science and Technology

The Web Page has various topics and many information. However, in many cases user demands only one part of Web Page. We suggest a search system showing a part domain of the whole page in the case of a Web Page search I read part of the page inside that a user demands in this study effectively, and to discover it. Materials choosing a Web Page increase, and it is hoped that a burden of the information discovery of the user is reduced by performing the presentation that assumed a domain a unit.

1. はじめに

近年，ひとつのページ中に多くの情報が記述されたWebページが増加している．例えば掲示板のスレッドやBlogのページおいてはページ中に複数の話題が記述され，様々な情報が混在し，多くの情報がひとつのページ中に記述されている場合がある．単一の話題のみ扱ったページであっても，Wikipediaのように詳細に記述されているページでは，非常に長い文章で記述されている場合や，箇条書きを多用するなどの記述で多くの情報が存在する場合がある．

このように多くの情報がページ中に記述されていても，利用者が目的とする情報はその全てではなく，ページ中の一部分のみである場合が多い．そのため，ページ中に目的の情報があるかどうかも含めて，必要な情報を得るには，利用者はページの選択や，ページ内での画面スクロールなどを繰り返して，情報が記述されている領域を発見する作業が必要となる．

筆者らは，Webページの部分領域を単位とした利用者への提示が重要であると考え，Webページを提示に適切なサイズに分割する研究を行っている[1]．分割に関する利用者実験の中で，部分領域提示には300~ 500字程度での分割が適切であるとの知見を得ている．これは検索エンジンの提示するスニペットよりも長く，ページを取捨選択するにはより多くの情報を提示すべきと考える．本論文では，利用者が求めると考えられる部分領域を検索の際に提示する検索システムを提案し，Webページの取捨選択を支援する．

2. 関連研究

2.1 Webページ検索支援に関する研究

Webページのタイトル・スニペット・URLなどの情報だけではページの取捨選択に十分とは言えず，求める情報の発見までに時間を要することから，検索結果のページ情報を効果的に提示する研究が盛んに行われている．

林ら[2]は，検索エンジンの示す検索結果に，特徴語を数件付随させて提示するシステムを提案している．検索結果として得られたそれぞれのページから特徴語を抽出して，上位5件を付随させることにより，利用者の要求を満たすページ到達の手間を減らす支援を行っている．

砂山ら[3]は，検索者の文脈やWebページ内容の把握を支援するためのHTMLテキスト分割システムを提案している．ページ中の検索語を含む前後の，意味のつながりのあるテキストを連結させることで，ページの文脈や内容把握の支援を行っている．

2.2 Webページ分割に関する研究

HTMLの繰り返し構造を利用して，Webページを分割・構造化する研究に南野ら[4]がある．Webページ作成者は，閲覧者にページのセグメントが分かるように記述する傾向がある．例えば同じタイプの項目が複数存在する場合は，各項目を同じように文字サイズや文字色で表現する．このような表現上の特徴が，HTML中の繰り返し構造に反映される傾向に基づいてページを分割する手法である．

ページ中の主要な部分をコンテンツ部分と称して検出する研究に吉田ら[5]，中村ら[6]がある．例えばニュース記事であれば，その本文部分をコンテンツ部分として定義し，ページ中からこのコンテンツ部分，もしくは非コンテンツ部分を検出する．

3. Webページのブロック分割

利用者が求めるページ中の一部分を効率的に閲覧・発見するために，Webページ検索の際にページ中の部分領域を提示する検索システムを提案する．システムの概要を図1に示す．システムは利用者から与えられた検索語を検索エンジンに渡して，結果の一覧を取得し利用者に提示する．利用者は提示された結果の一覧から，いずれかのページを指示する．システムは指示されたページのスニペット周辺部分を，レイアウト構造を保持したままのブロックとして取得し，付随させ提示する．この部分領域の付随には，大きく「Webページ分割」と「部分領域提示」の2つの処理が必要となる．本章ではWebページ分割について述べる．

連絡先: 田崎 雄一郎，筑波大学大学院 図書館情報メディア研究科 佐藤研究室，〒305-8550 茨城県つくば市春日 1-2，tasaki@slis.tsukuba.ac.jp

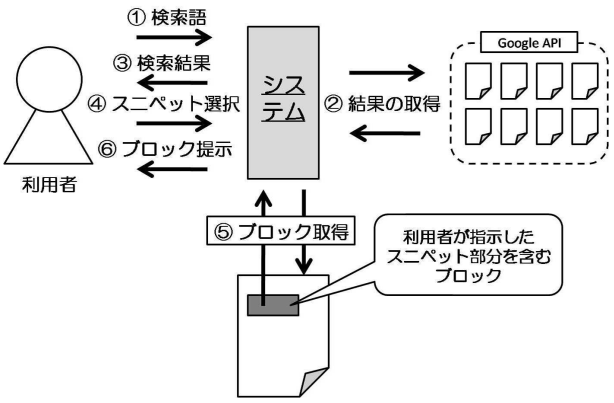

図 1: システムの概要

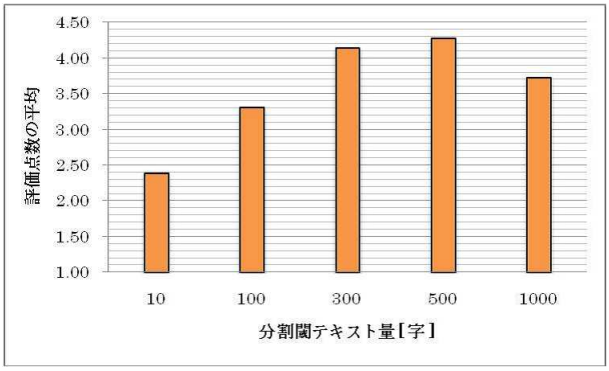

図 2: 分割閾値別の平均評価

3.1 Web ページの部分領域分割

前章で述べたように，Web ページ分割やコンテンツ抽出に関する研究は盛んに研究されているが，これらの手法は提示に適したサイズを考慮していない．また，利用者はコンテンツ部分のみを求めるとは限らない．本論文では利用者の提示に Web ページ分割手法を提案する．

3.1.1 構造的な切れ目による分割

Web ページは一般に Hyper Text Markup Language（以下，HTML）によって記述される．HTML はタグによって要素を構成する形式で記述され，タグは大きく「ブロックレベル要素」と「インライン要素」に分けられる．HTML の厳密な規定では，body タグの直下にはブロックレベル要素しか記述することができないため，これにより大きく構造の切れ目が表わされると考えられる．

そこで，HTML のブロックレベル要素を主に利用し，そこにいくつかのタグを追加・除外することとした．追加したタグには，例えば table タグがある．table タグは本来表の作成に用いらインライン要素であるが，Web ページのレイアウト作成にも用いられることが多いため、構造の切れ目として有用なタグであるとした．

表 1: 構造的な切れ目として扱うタグ

address, blockquote, div, dl, fieldset, form, hr, ol, p, pre, table, td, tr, ul

3.1.2 ブロックの細分化と結合

前節で述べた HTML タグに基づく分割ではブロックのサイズ，すなわち個々のブロックに含まれる情報の分量を考慮していない．このため，ひとつのブロックが多くの情報を持つ場合，逆に情報の分量が少なすぎる場合には，ブロックの大きさが提示に適したものとなっていない可能性があり，部分領域提示に適した分割を実現できない．そこで，ブロックの細分化と結合を行うことで提示に適したブロックを決定する手法を提案する．

ブロックがどの程度情報の量を持つかどうかはブロックのテキスト量で判断し，ある一定のテキスト量を閾値として定め，その閾値を超えるブロックを細分化し，閾値に満たないブロック同士を結合する手法である．

ブロックの細分化は，HTML の改行要素を用いる．具体的には br タグや li タグである．li タグはそれぞれで分割し，br タグはブロック内で連続する数の平均値以上の部分でのみ分割することとした．平均値以上での分割は，改行を多く用いて読みやすいレイアウトを作成しているページへの対応を考慮したものである．

表 2: 閾値別の平均評価

	B-1	B-2	W-1	W-2	N-1	N-2
10 字	3.33	2.83	1.33	1.50	2.33	3.00
100 字	4.00	4.17	3.00	2.67	2.83	3.17
300 字	**4.17**	**5.17**	3.83	**4.17**	3.17	4.33
500 字	3.67	3.67	**4.00**	4.00	**5.00**	**5.33**
1000 字	3.33	2.50	2.83	4.17	4.83	4.67

ブロック同士の結合は，構造的な関係を維持するため，HTML の記述順と階層構造を基に行った．単に HTML 記述順にブロックのテキスト量で結合すると，階層構造が離れたブロック同士で結合される場合がある．本論文ではブロックの階層構造における親ノード，もしくは兄弟ノードのみを対象として，あらかじめ定めたテキスト量の閾値を超えるまで繰り返し結合した．

3.2 ページ分割に関する利用者実験

3.2.1 適切なブロックサイズの調査

提案手法において適切なブロックサイズを明らかにする目的で利用者実験による予備調査を行った．実験は提案手法を用いて実際に Web ページ分割を行い，その結果の評価を男女 6 名の実験参加者に依頼した．実験の対象ページ・分割段階などは以下の通りである．

- 評価対象のページは，Blog 記事 (B-1，B-2)，Wikipedia 記事 (W-1，W-2)，ニュース記事 (N-1，N-2) をそれぞれ 2 件ずつ，計 6 件[*1]
- 分割の閾値は，10・100・300・500・1000 字の 5 段階

被験者には，予め本論文の目指す部分領域の提示について説明を行い，「部分領域提示に適した分割かどうか」を特に注意して評価を依頼した．

*1 B-1 http://blog.livedoor.jp/dqnplus/
B-2 http://kanteiblog.typepad.jp/
W-1 http://ja.wikipedia.org/wiki/図書館情報大学
W-2 http://ja.wikipedia.org/wiki/つくばクレオスクエア
N-1 http://www.yomiuri.co.jp/national/news/20100117-OYT1T00029.htm
N-2 http://www.yomiuri.co.jp/entertainment/news/20100116-OYT1T00261.htm

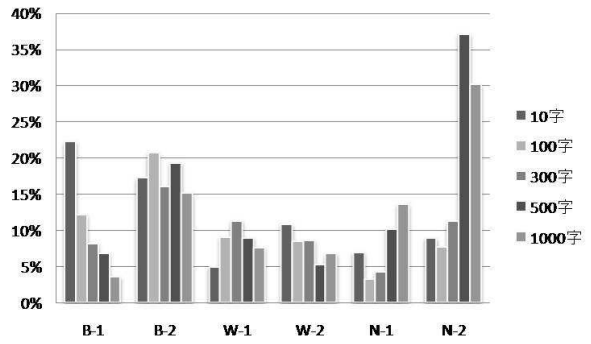

図 3: 被験者による分割ブロックの再現率

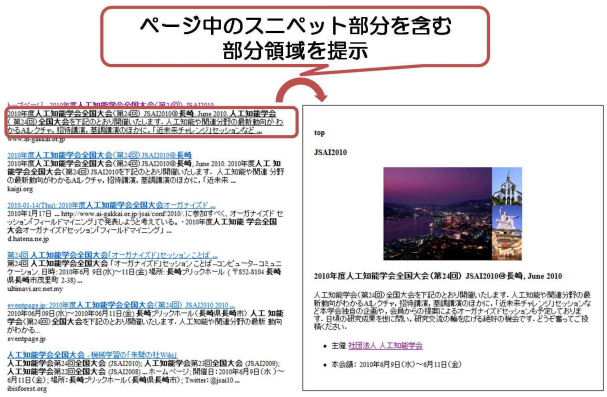

図 4: システムの実行例

利用者実験の結果を表 2 に示す．分割の適切さは 1~ 6 点までの 6 段階で評価し，点数が高いほど適切な分割との評価を得ている．図 2 は分割閾値別に，利用者による評価の平均を表したものである．この結果から，本論文で提案手法を用いて 300 字，もしくは 500 字の閾値で分割を行った場合に評価点 4.0 以上の評価を得ていることが分かる．この評価点 4.0 は実験協力者に対し,「どちらかと言えば適切な分割」と指標した点数である．同時に表 2 より，実験対象としたそれぞれのページは全て 300 字，もしくは 500 字のときに最も高い評価を得ていることが分かる．これらのことから，本手法を用いて 300 ~ 500 字でページの分割を行えば，提示に適したであることが明らかとなった．

この結果を受けて，まず前節までに述べた手法における分割・結合のテキスト量閾値を 300 字とした．その後，200 字以下のブロックをさらに結合させやすくすることで，300~ 500 字の閾値として結合を行った．200 字以下のブロックの結合は，親ノード・兄弟ノードのみを結合対象としていたものを，さらに段階を増やしたものまで結合の対象とすることで拡張した．これは結合にふさわしい対象がない場合に，テキスト量が少ないブロックでもそのまま保持されてしまうことを避ける狙いがある．

3.2.2 被験者分割ブロックの再現率

利用者実験において，被験者に紙面上に印刷した Web ページを提示し，そこに手動で線を書き込むことで，実験協力者自身が思う適切なブロック分割も依頼した．依頼した Web ページの分割結果には，一部分割が行われなかった箇所があるなど，いくつか利用できないデータも存在した．それらのデータは集計の対象から除外した．

システムが分割したブロックを正解として，実験協力者が分割したブロック中にいくつ正解が含まれるかを集計した．このとき，一行でもブロックの分割位置が異なれば，不正解とした．システムが分割した正解のブロック数も集計し，実験協力者の分割結果がどの程度正解ブロックを再現できているかを図 3 に示す．ここで再現率は以下の式で与える．

$$\text{再現率} = \frac{\text{実験協力者が分割したブロック中の正解数}}{\text{実験協力者が分割したブロックの総数}}$$

実験協力者が分割したブロックと一致するものは，システムが分割した正解ブロック中には平均で 15 %程度であることが分かった．本評価では利用者が分割したものと一行でも分割位置が異なれば不正解としたことも，低い再現率になった要因のひとつであると考えられる．

4. 部分領域を提示する検索システム

一般的な Web ページ検索エンジンにおいて，利用者が検索語を入力すると、それに対したページのタイトル，スニペット，ページの URL などが一覧で表示される．利用者は提示された一覧の結果から，自身が求める情報，ひいてはその情報が含まれるページを繰り返し取捨選択する必要がある．その際ページ内に検索語が含まれる部分であるスニペットなどを基に取捨選択を繰り返すが，これらが十分に選択のための情報を持っているとは言えないと考える．前節の部分提示に適したブロックに関する利用者実験においても，300~ 500 字程度のテキスト量が部分提示に適しているとの結果を得ており，これは通常 120 字程度で記述されるスニペットよりも明らかに多いものである．このため利用者は検索エンジンから一覧で提示される結果が取捨選択に十分ではないと感じていると言える．そこで本論文では，Web ページ検索の際にページ中の部分領域をスニペット部分に付随して提示する検索システムを提案する．

4.1 システムの構築

Web ページ検索結果の取得には GoogleAPI[*2] を用い，利用者が入力した検索語に関連のある Web ページ一覧を提示する．利用者は提示された結果の一覧から，目的の情報が含まれるページの選択を繰り返す．このとき本システムは，利用者によって指示された検索結果のスニペット部分を，図 4[*3] のように画面右側に，そのスニペット部分を含むページ内のブロックを提示する．部分領域提示に際して画面内の提示領域は限られているので，ブロック内のテキストのフォントサイズと画像の表示サイズは大きさを限定して提示している．

4.2 システム評価

本システムでは，Web ページ内からスニペット部分を検出するために PHP の similar_text 関数を用いて，スニペットと最も類似度の高い文が含まれるブロックを検出している．この手法により，どの程度の精度でスニペットが含まれるブロックが検出できるかを検証した．検証に用いた検索語は以下の通りである．

- 検索語は，広辞苑に掲載された見出し語から無作為に抽出
- 4 通りの検索語

*2 Google AJAX Search API: http://code.google.com/intl/ja-JP/
*3 http://www.ai-gakkai.or.jp/jsai/conf/2010/

- 1 つの検索語に対して、上位 40 件の結果

これらの検索語を用いて検証した結果を表に示す.

表 3: 利用者分割ブロックのシステム再現率

	文の数	正解率
複数に分割されたスニペットを構成する数	228	0.469
非分割対象を除外後	125	0.856

検索結果件数は 160 件であるが，スニペットは複数に分かれたものも存在する．それらを考慮したスニペットを構成する文の数は，のべ 228 文であった．その中でページ中からスニペットを含む領域を検出できたのは，107 文であった．これは表 3 に示す通り約 47 %の精度であり，十分な精度で検出できたとは言えない．正しく検出できなかったページは，スニペットが画面に表示されない形式であるヘッダ中に記述されている場合などがあった．本手法では similar_text 関数を用いて，スニペットと最も類似度の高い領域を検出している．そのためスニペットがヘッダ中に含まれて画面中に表示することができずとも，最も類似の領域を取得し表示しているため問題ないと考える．他の検出できない場合に，検索結果が PDF や Word ファイルなどであった場合ある．本論文では HTML の階層構造などに基づいてブロック分割を行っているため，PDF など HTML 以外の形式は分割の対象としていない．このことから，以上のような結果を除外して精度を再測定した．このとき除外後のスニペットを構成する文の数は 125 文となり，正解数は 107 文と変わらないため，スニペットを含む領域の検出精度は約 86 %となる．

4.3 ブロックの分割位置に関する考察

本システムでは，ブロックがどの程度利用者が満足な範囲で取得されているかが重要である．前章で実験協力者が分割した結果の再現率を示したが，平均で約 15 %程度であり，一致したブロック数はあまり多くなかった．しかし，実験協力者による分割がシステムの結果と完全に一致しておらずとも，十分な提示が行えると考える．例えば図 5[*4]の結果 A と結果 B のようにブロックの切れ目が異なっていたとしても，その切れ目の部分に意味があるわけではないため，部分領域の提示に際して問題ないと考える．

しかし，それでは純粋にスニペット周辺を文字列で切り出すのみでも構わないのではないかという議論が考えられる．そのため，今後は我々が提案している分割手法が，単純にスニペット周辺を切り出すだけの手法に比べた有用性を示す利用者実験を行う予定である．この実験により，本研究が階層構造で分割・結合対象を限定していることによる，純粋なテキスト量の身による抽出より有用な結果が得られることが期待される．

5. おわりに

本論文では，検索エンジンの示す検索結果に，それぞれの Web ページ中の部分領域を提示する検索システムを提案した．部分領域提示のために，多くの情報が記述された Web ページの一部分を，ページのレイアウトを保持しつつ，部分領域提示に適した分量の情報を持ったブロックに分割する手法も提案・実装した．

ブロック分割に関する利用者実験を行い，本手法を用いたブロック分割において，300~ 500 字のテキスト量が提示に適

*4 http://ce.slis.tsukuba.ac.jp/

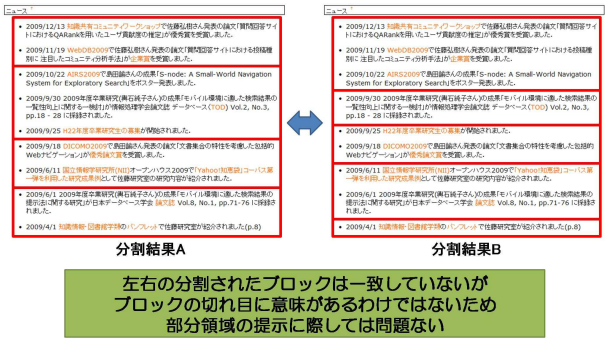

図 5: 相違するブロック分割位置

切な大きさとなることが分かった．適切な大きさに分割されたブロックを用いて，Web ページ検索の際にページ中の部分領域をスニペット部分に付随して提示する検索システムを構築した．

システムは対象とした HTML 形式のファイルに対して，十分な精度でスニペット周辺のブロックを取得することができた．検索結果へのブロックの付随が Web ページ検索において有用であることの検証が今後の課題である．

謝辞

本研究の一部は科研費 (21500091) の助成を受けたものである．ここに記して謝意を示す.

参考文献

[1] 田崎雄一郎，佐藤哲司：Web ページの階層的な分割と提示に関する一検討，第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010)，A4-2，2010.

[2] 林祐平，品川徳秀：特徴語付きウェブ検索インタフェースの提案，電子情報通信学会 第 19 回データ工学ワークショップ (DEWS2008)，B5-6，2008.

[3] 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.87，No.12，pp.1089-1097, 2004.

[4] 南野朋之，齋藤豪，奥村学：繰り返し構造を用いた web ページの構造化に関する研究，情報処理学会研究報告，自然言語処理研究会報告，Vol.2003，No.23，pp.185-192, 2003.

[5] 吉田光男，山本幹雄：教師情報を必要としない Web ページ群の主要コンテンツ自動抽出，第 23 回人工知能学会全国大会 (JSAI2009)，2B3-1，2009.

[6] 中村達也, 白井清昭：ウェブページにおける非コンテンツ領域の検出，言語処理学会年次大会発表論文集, vol.13, pp.234-237, 2007.