

# Web ページの階層的な分割手法と提示に関する一検討

田崎雄一郎<sup>†</sup> 佐藤 哲司<sup>††</sup>

<sup>†</sup> 筑波大学図書館情報専門学群 〒 305-0821 茨城県つくば市春日 1-2

<sup>††</sup> 筑波大学大学院図書館情報メディア研究科 〒 305-0821 茨城県つくば市春日 1-2

E-mail: <sup>†</sup>s0813179@u.tsukuba.ac.jp, <sup>††</sup>satoh@slis.stukuba.ac.jp

**あらまし** ニュースサイトのトップページやブログなどの Web ページには様々な情報が混在しているが、利用者が求める情報はその中の一部分だけである場合が多い。また、利用者が求める情報がページ中に含まれるかどうかの取捨選択を行うためには、判断に必要かつ十分な範囲で Web ページの一部分を効率的に閲覧することが必要である。本論文では、Web ページの構造的な切れ目を用いて Web ページを分割し、さらにテキストの分量を用いて分割・結合することで作成されたブロックを単位として提示する手法を提案する。分割手法について利用者実験を行い、本手法においてどの程度のテキスト量を用いれば提示に適切な Web ページ分割が行えるかを検討した。

**キーワード** Web ページ分割, 部分提示, 情報選択

## A Study of Hierarchical Division and Partial Presentation of Web Pages

Yuichiro TASAKI<sup>†</sup> and Tetsuji SATOH<sup>††</sup>

<sup>†</sup> School of Library and Information Science, University of Tsukuba

Kasuga 1-2, Tsukuba, Ibaraki, 305-8550 Japan

<sup>††</sup> Graduate School of Library Information and Media Studies, University of Tsukuba

Kasuga 1-2, Tsukuba, Ibaraki, 305-8550 Japan

E-mail: <sup>†</sup>s0813179@u.tsukuba.ac.jp, <sup>††</sup>satoh@slis.stukuba.ac.jp

**Abstract** Various information exists together on the Web Page such as the top page of news site and blog, etc. However, in many cases, the user demands information that only part of the Web page. The user needs to effectively reading the part of Web page that is enough to need for judgment the Web page include the demand information. This article suggests the partial presentation technique that assumed the block as a unit. The block is made by Web page division. The Web page is divided with structural break of the page, and based on the quantity of the texts in the page. We performed a user experiment about Web page division technique and examined the appropriate partial presentation if we used quantity of how much text in this technique.

**Key words** Web page division, partial presentation, information select

### 1. はじめに

近年、ひとつのページに様々な情報が記述された Web ページが増加している。この背景には、単にインターネット上における情報量が増加していることも一因であるが [1], [2], 利用者の誰でもが情報を発信できる環境が多様化したことが大きいと考える。利用者の容易な情報発信の手段として、掲示板, Blog, SNS やつぶやきサイト, 口コミサイトなどの CGM コンテンツや、簡単に Web ページの編集が可能な Wiki コンテンツなどがある。例えば, Blog は月に 40 万から 50 万件の新規開設があることから [2], 利用者により作成されたコンテンツが増加していると言える。

このように利用者によるコンテンツにおいては、複数の話題が記述されるページが多い。例えば掲示板のスレッドや Blog のページにおいては、ひとつのページ内に複数の話題が記述され、ページ中に様々な情報が混在している。このことにより、単一ページあたりのテキスト記述量が増加していると考えられる。

また単一の話題について書かれた Web ページであっても、その話題について詳細に書かれていると、ページ中のテキスト記述量は増加する。例えば Wikipedia のように詳しく事象を解説したページにおいては、詳細な解説のために非常に長い文章で記述せざるを得なかったり、箇条書きが多用されたりと、ひとつのページ内に非常に多くの情報が含まれている。

このように様々な多くの情報がページ中に含まれていたとし

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、利用者は画面の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。

本論文は、様々な情報が混在する Web ページ中からの、利用者の情報発見の負担を軽減することを目的とする。そのためには Web ページ中の全ての情報ではなく、利用者が目的とするページ内の領域を容易に閲覧できることが必要であると考えられる。本論文では多くの情報が含まれる Web ページの一部分を、ページのレイアウト情報を保持しつつ、あらかじめ指定した分量の情報をを持ったブロックを単位に分割し提示する手法を提案する。

本論文の構成を以下に示す。2. で先行研究について述べ、本研究の位置付けを明確にする。3. で提案する Web ページの分割方法について述べ、さらにその提示案についても触れる。4. で分割手法を評価し、5. で考察と議論を行う。最後に 6. で本論文をまとめる。

## 2. 先行研究

本論文で提案する内容は、ブロックを単位とした Web ページの分割手法とブロックを単位とした提示手法とからなる。2.1 で Web ページの分割に関する先行研究を、2.2 で Web ページの部分提示に関する先行研究を説明する。

### 2.1 Web ページ分割に関する先行研究

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 [3] 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に注目してページの一部を抽出する手法も盛んに研究されている [4,5]。例えばニュース記事であれば、その本文部分を主要な部分として定義し、ページの中から主要な部分をコンテンツ部分と称して抽出する。またこの主要な部分をコンテンツ部分と呼ぶものも多い。

吉田ら [4] は、Web ページ中からの主要部分を抽出している。Web ページのコンテンツ部分は他の Web ページには出現せず、ひとつの Web ページ中にのみ出現する傾向がある。一方、複数のページにまたがり何度も出現する部分は、コンテンツ部分ではないという傾向を用いて主要部分を抽出している。

中村ら [5] は Web ページ中の非コンテンツ領域を検出している。例えば広告、アクセスカウンタ、検索フォームなどを非コンテンツ領域であると定義し、ページ内からそれらを検出する手法について述べている。具体的には、非コンテンツ部分を示唆するキーワード、テキスト長、内部リンクと外部リンクの含有などによって非コンテンツ領域を検出している。

本論文では、提示を行うための分割であるから、分割後の情報の分量が重要となる。これに対して従来の分割手法は、分割

後の情報の分量が膨大または極小で、提示に適さない大きさとなっている恐れがある。また、掲示板や Blog のようにコンテンツ部分がはっきりとしない、あるいは、コンテンツ部分が複数存在するページも多く存在する。本研究の着眼点は、Web ページ中に利用者が必要とする情報とそうでない情報が混在しているという点で、従来研究と共通するところがある。しかし利用者は必ずしもコンテンツ部分の情報を求めているとは限らないと考えるため、本論文で用いる部分抽出に、これらの方法を直接適用することはできない。

### 2.2 部分提示に関する先行研究

Web ページの部分提示を行う研究には、モバイル端末を対象としたものが多く、PDA 端末に適した部分提示法の研究に、山本ら [6]、Chen ら [7] がある。山本らは、HTML を特定のタグによって分割した後に、利用者の履歴情報によって学習を行い、必要な部分のみを提示している。Chen らはページ全体をヘッダーやフッターなどに分割し、さらに HTML タグを用いて分割している。画面上に縮小されたページ全体を提示し、選択された部分を拡大するなどの手法で PDA 端末への適応を目指している。

これらの研究は、モバイル端末の小さい画面では PC 向けに作成された大きな Web ページの閲覧が困難であることを改善するものである。しかし近年では PC で閲覧するにも情報量が多すぎて、閲覧が容易には行えないページが増えている。そのため本研究では PC 向けの支援を行う手法を提案する。

Web ページを分割、あるいは部分を抽出して提示する研究も盛んに行われている [8,9]。韓ら [8] は HTML をの木構造に変換し、ページの分割を行っている。レイアウト構造の類似した 2 つのページから部分抽出を行い、両方を同一のページで提示する提案をしている。砂山ら [9] は、ブロックレベル要素、複数のリンク、句点が存在する個所で Web ページの分割を行い、部分ごとのテキストをセグメントに分割している。分割したセグメントを、外部の重要文抽出システムを通して得られた重要文の提示を提案している。

これらの研究は、韓らの研究では情報の分量が考慮されていない分割であること、想定する提示方法が全く違うことから、本研究とは目的・手法ともに異なる。また砂山らの研究では、部分中に含まれる情報の分量は考慮されているが、レイアウト構造を保持した提示を行っていない。また、情報の分量がスニペットに適したものに限定されており、本研究で提案するテキストの分量を考慮した分割とはなっていない。

## 3. Web ページ分割法の提案

### 3.1 Web ページ分割の概要

本論文では、レイアウト情報を保持したまま抽出したページの一部をブロックと呼ぶ。ブロックを単位として提示を行うための第一フェーズの処理であるブロック分割について述べる。

本論文による Web ページのブロック分割は、提示に適した分割を目的としているため、ページ中の内容は考慮されていない。これは内容の区切りで分割を行うと、サイズが大きすぎる、もしくは小さすぎて提示に適さない場合が多く存在するからであ



図 1 Web ページ分割の流れ

る。分割の手順は、まず入力された Web ページに対して、コメント部分を除去するなどの前処理を行う。その後、Web ページを構造的な切れ目で分割し、情報の分量を指標とすることで提示に適切な大きさのブロックとする。本章では図 1 に示した処理の流れに沿い、まず 3.2 で Web ページの構造的な切れ目によって構造を再構成する方法について述べる。3.3 で情報の分量を用いた細分化を行い、最後に 3.4 である程度の情報を保持するようブロック同士の結合を行う方法について述べる。

### 3.2 Web ページの構造的な切れ目による分割

Web ページは一般に、HyperText Markup Language（以下、HTML）を用いて記述される。HTML は World Wide Web Consortium<sup>(注1)</sup>（以下、W3C）によって規格が策定されている。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。W3C の規定によると、body タグの直下には厳密には表 1 に示す、ブロック要素のみしか記述することができない。そのため、これらのブロックレベル要素により、大きく構造の切れ目が表わされると考えられる。

表 1 に示したブロックレベル要素の中から「h1 ~ h6」、「center」、「dir」、「isindex」、「menu」、「noframes」を除き、「tr」、「td」タグを追加した、表 2 に示すタグを本研究では Web ページの構造の切れ目を表すものとして扱う。

これらのタグを追加・除外した理由を述べる。h1 ~ h6 タグは確かに構造の切れ目としての役割を持つが、これらのタグは一般に見出し要素として利用される。このため h1 ~ h6 要素内に記述される内容は、他のブロック要素のタグが持つ情報の分量より、非常に少ない場合がほとんどである。情報の分量が少ない場合は、後述するブロックの結合により適切に結合されるが、階層構造が離れてしまう場合にはこれが難しくなる。構造の切れ目で分割を行うと、その部分で階層が再構成される。そのため、情報の少ないタグで構造を切ると、階層が離れて結合が困難になることから、階層を持たせることは望ましくない。このことから、これらのタグは本研究において明示的な分割の対象としない。

また、表 1 に示したタグは、W3C によって規定された HTML の移行型（Transitional）である。HTML の規定には移行型と

表 1 HTML 規定別のブロックレベル要素（移行型）

タグの種類	推奨（厳密型）	非推奨
見出しを表すタグ	h1, h2, h3, h4, h5, h6	
箇条書きを表すタグ	ul, ol, dl	dir, menu
特定の領域を表すタグ	address, blockquote	
入力フォーム関連のタグ	form, fieldset	isindex
レイアウト関連のタグ	pre	center
段落を表すタグ	p	
表を作成するタグ	table	
水平線を描画するタグ	hr	
フレーム関連のタグ		noframes
特に意味のないタグ	div	

表 2 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
---

厳密型（Strict）があり、移行型に含まれるタグは、厳密型で定義されるタグでは非推奨のものもある。本研究では厳密型で非推奨と規定されるタグについては、構造の切れ目として利用しない。

構造の切れ目を表すとして追加した td, tr は、本来テーブルの要素を定義するために用いられる。しかし本来の用いられ方ではなく、内容を二分することや枠線の表示が容易であることから、ページレイアウトを整えるために用いられることも多々ある。このことから、Web ページ作者が構造の切れ目として td, tr タグを用いていることが想定される。そのため本研究では、これらのタグも構造的な切れ目として扱う。

以上のことから表 2 に示すタグを本研究では構造的な切れ目を表すタグとして扱い、その部分で構造の再構成を行った。

W3C の規定では、body タグの直下にはブロック要素しか記述できないと上記したが、全ての Web ページがこれに沿って厳密に作成されているわけではない。また、本研究ではブロック要素であっても、一部のブロックは構造的な切れ目として採用しなかった。そのため、構造的な切れ目として扱った特定のタグで構造的に分割した部分をブロックとしただけでは、どのブロックにも属さない部分が生じる。そのため、それらの部分を包括的にブロック化する必要がある。

本研究では、構造的な切れ目で分割されたブロックの隙間それぞれをひとつのブロックとみなし、HTML 文書を再構成した。これにより HTML 記述中の全部分をブロックとして扱うことができる。

### 3.3 ブロックの細分化と結合

前節で述べた切れ目で分割しただけでは、HTML タグに基づく分割に過ぎないので、その内部の情報の分量は考慮していない。そのため、ひとつのブロックが多くの量の情報を持つ場合や、その逆の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る。本論文では Web ページの部分提示を行うことを目的としているので、分割されたブロックの細分化と結合を行う必要がある。

ブロックがある程度の量を持っているかどうかはブロックの

(注1) : <http://www.w3.org/TR/html4/>

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。

本研究では、改行要素を用いて細分化を行う。具体的には `br` タグと `li` タグを用いる。`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した。これは、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に対応するためである。このようなページにおいて、単純に改行でひとつずつ分割したのでは、それぞれのブロックが非常に細かいものになってしまう。そのため連続する改行の平均以上の部分で分割することで、ある程度構造のまとまった分割ができると考えた。また、改行要素が用いられていないテキスト量の大きいブロックに関しては、句読点による分割が考えられる。しかし改行を用いず記述するということは、ページ作者がそれでひとつのまとまりを持たせていると考えられることから、無理に分割する必要がないと判断した。

分割されたブロックは、反対にテキスト量が少な過ぎる場合もある。改行タグを単位として分割を行ったため過剰に細分化された場合や、そもそも構造的な意味の切れ目として用いた要素の段階でのテキスト量が少なすぎた場合などである。このような場合、部分提示に用いるブロックとして適切であるとは言えない。そのため、ブロックがある程度のテキスト量を持つように、ブロック同士を結合する。

結合の対象となるブロックは、HTML の記述順が前後であるブロックかつ、図 3 のように表わされる階層構造において、自身の兄弟ノードとなるブロックか、自身の親ノードとなるブロックのみである。これは、HTML 記述では隣接して記述されているも構造的な関係性がない場合に、階層構造によって結合を防ぐことを目的としている。

例えば図 2 の HTML 記述は、本研究では図 3 のように再構成される。このとき、図 2 において、「セル 10」のブロックは HTML 記述では「テキスト 2」と「セル 11」のブロックと隣接している。しかし図 3 の階層構造においては「セル 11」とは兄弟ノードであるものの、「テキスト 2」とは親子ノードでも兄弟ノードでもなく、階層構造が離れている。このように HTML 記述で隣接していても階層構造が離れている場合は、ブロック結合の対象としない。

結合の際も、ある一定のテキスト量を閾値として定め、閾値に満たないブロックを別のブロックと結合させる。上記の通り階層構造に基づき結合するので、テキスト量が少ないブロックであっても、結合にふさわしいブロックが存在しなければ、結合を行わずそのままの状態でも保持される。

### 3.4 ブロック分割の実行例

分割の実行結果の一例を図 4 に示す。図 4 は「筑波大学 図書館情報専門学群」の Web ページを、作成したシステムで分割したものである。このとき、分割と結合に用いるテキスト量の閾値は 500 字である。図中では赤枠によってブロックが区切られているが、実際の実行結果は HTML 記述部分が得られるのみであり、枠は結果を基に手動で描いたものである。

```
<body>
  テキスト 1
  <div>
    <ul>
      <li>リスト 1</li>
      <li>リスト 2</li>
    </ul>
    <p>文章</p>
  </div>
  テキスト 2
  <table>
    <tr>
      <td>
        セル 1 0
        セル 1 1
      </td>
      <td>セル 2</td>
    </tr>
  </table>
</body>
```

図 2 HTML のサンプル記述

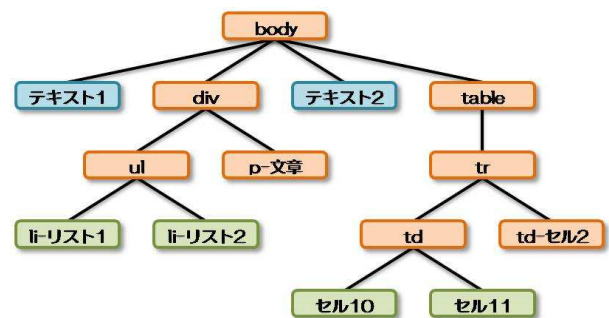


図 3 構造化されたサンプル HTML 記述

### 3.5 ブロックを単位とした提示案

ブロック単位に分割された Web ページの提示手法の応用例として、検索エンジンに付随させた提示手法を提案する。検索エンジンは利用者が入力した検索キーワードを基に、Web ページのタイトルと、スニペットを返す。スニペットとは Web ページ中の、利用者が入力した検索キーワードが含まれる一部のテキストである。スニペットは情報発見のための一助となるが、プレーンテキストでの記述であることや、情報の分量が少ないことから、結果の比較・判断に際して必要な情報が不足する場合がある。レイアウト情報が保持され、ある程度の情報を持ったブロックをスニペットの周辺情報として付随させることで、利用者の情報発見を支援することができると期待する。不足した情報を補うために、単純にスニペット周辺の情報を取得する方法も考えられるが、HTML 記述の前後関係などを考慮した本論文で生成されたブロックは、このような提示に適した分割となっていると考える。

## 4. ブロック分割に関する利用者実験

### 4.1 利用者実験の概要

前章で提案したブロック分割アルゴリズムを実装し、分割が有効に行えるかを利用者実験によって確認する。

実験では、被験者にシステムによって得られた Web ページの分割結果を提示し、適切な分割が成されているか評価しても





図 4 分割結果の一例



図 5 提示イメージ

表 3 分割段階

	分割・結合関 (文字)
分割 A	10
分割 B	100
分割 C	300
分割 D	500
分割 E	1000

表 4 評価点数の基準

評価点数	評価基準
1	不適切な分割である
2	あまり適切に分割されていない
3	どちらかと言えば適切ではない
4	どちらかと言えば適切である
5	ある程度適切に分割されている
6	適切に分割されている

対象に行った。実験は 6 つの Web ページ、5 つの分割結果を利用した。

6 つのページには Blog のトップページ、Wikipedia 記事、ニュース記事を各 2 つずつ用い、本論文ではそれぞれ Blog.1<sup>(注2)</sup>、Blog.2<sup>(注3)</sup>、Wiki.1<sup>(注4)</sup>、Wiki.2<sup>(注5)</sup>、News.1<sup>(注6)</sup>、News.2<sup>(注7)</sup>と記述する。各ページの情報記述量はほぼ同じになるものとした。閾値別の 5 つの分割結果は、それぞれのページを対象に表 3 に示す 5 段階で分割したものを提示し、表 4 に示す 6 段階で評価を依頼した。

実験は以下に示す手順を、6 つのページそれぞれに対して順番に実施する。

- (1) 被験者に分割の対象となる Web ページを印刷した紙面を配布する。
- (2) 紙面上に、被験者が考えるブロックの分割線を手書きで書き込む。
- (3) 表 3 で示した 5 種類の分割結果を提示し、分割が適切に分割されていると考えるかどうかを表 4 の 6 段階で評価する。
- (4) 評価の根拠があればコメントを記述する。  
(特に明確な理由がなければ記述しなくても構わない)
- (5) システムの提示する 5 つの分割結果を、適切だと思われる順に並べる。

本実験は特に時間制限を設けず行った。実験参加者には、あらかじめ本論文が目指すブロック単位での提示について説明し、分割が「提示に適した大きさで行われているかどうか」を特に注意して分割と評価を行ってもらった。この注意は、被験者はページレイアウトに沿った切れ目でのみ分割する可能性があり、それだけでは提示するには大きいサイズである場合に、その内部でも分割してもらうことを想定している。

らう。これを分割・結合に用いるテキスト文字数の閾値を変えた複数の結果で提示することで、どの程度のテキスト量で提示に適切なページ分割が行えるか評価する。

## 4.2 利用者実験の方法

本実験は、図書館情報学を専攻する大学 4 年生の男女 6 人を

(注2) : <http://blog.livedoor.jp/dqnplus/>

(注3) : <http://kanteiblog.typepad.jp/>

(注4) : <http://ja.wikipedia.org/wiki/図書館情報大学>

(注5) : <http://ja.wikipedia.org/wiki/つくばクレオスクエア>

(注6) : <http://www.yomiuri.co.jp/national/news/20100117-OYT1T00029.htm>

(注7) : <http://www.yomiuri.co.jp/entertainment/news/20100116-OYT1T00261.htm>

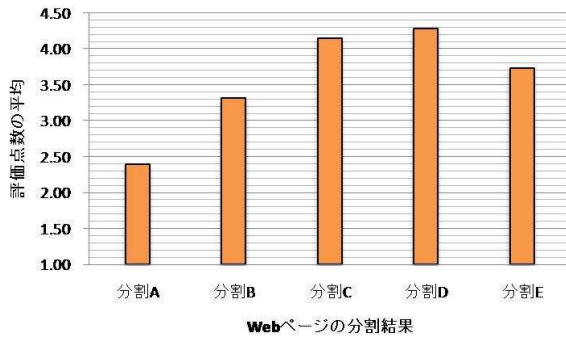


図 6 分割結果ごとの利用者による平均評価

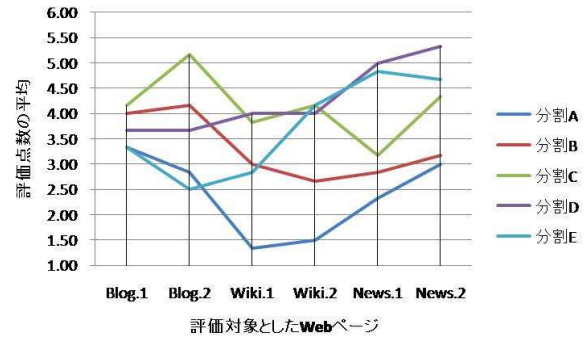


図 7 Web ページごとの利用者による平均評価

6 つの対象ページ全ての実験を終えた段階で、実験手順 (2) における分割の際、6 つの内どのページの分割が最も困難であったか、最も容易であったかとその理由についてアンケートをとった。

#### 4.3 利用者実験の結果と評価

本実験によって得られた、結果ごとに被験者の評価値の平均値をまとめたものを図に示す。またこの平均値のデータから、分割結果ごとの平均評価に着目した図 6 と、Web ページごとの平均評価に着目した図 7 を示す。

表 5 ページ、結果ごとの平均

	Blog.1	Blog.2	Wiki.1	Wiki.2	News.1	News.2
結果 A	3.33	2.83	1.33	1.50	2.33	3.00
結果 B	4.00	4.17	3.00	2.67	2.83	3.17
結果 C	4.17	5.17	3.83	4.17	3.17	4.33
結果 D	3.67	3.67	4.00	4.00	5.00	5.33
結果 E	3.33	2.50	2.83	4.17	4.83	4.67

#### Web ページ分割に最適なテキスト量

図 6 は、横軸が表 3 で示した A から E の 5 種類の分割結果、縦軸が 5 種類の結果それぞれにより分割されたページの、全被験者の平均点数である。この図より、結果 A から E のいずれが、対象となった全ページの平均として高い評価を得たかを確認でき、結果 D のとき、次いで結果 C のときに高い評価点を得ていることが分かる。

#### 最適なテキスト量のページ種別依存

図 7 は、横軸が対象となった 6 つの Web ページ、縦軸がページの分割結果 A から E の、それぞれにおける被験者による評価の平均値である。この図より、提示に適切なサイズがページの種別に依存するかどうかを確認でき、対象となった 6 つの Web ページすべてにおいて、結果 C、D のときに最もよい評価を得ていることが分かる。

## 5. 考 察

本論文では、多くの情報が含まれる Web ページ中の一部分を、ブロックを単位として提示することを提案した。ブロックを提示するために、Web ページをブロックに分割する手法を考案し、どの程度の大きさの分割が提示に適切なものか調査する利用者実験を行った。本章では、前章の実験結果を踏まえ考察

する。

#### Web ページ分割に最適なテキスト量

図 6 から、分割結果 D、次いで結果 C のときに評価が高いことが分かる。さらに評価結果を示した図 7 から、Blog.1、Blog.2、Wiki.2 では結果 C のときに最も適切に分割され、Wiki.1、News.1、News.2 では結果 D のときに最も適切であり、すべてのページにおいて結果 C、または結果 D が最も良い評価を得ている。この最も良い評価は、対象となった 6 つのページ全てにおいて「どちらかと言えば適切な分割」である 4.0 以上となっている。このことから、テキスト文字数 300~ 500 字で分割されたブロックが、部分提示に適切なサイズであると言える。この 300~ 500 字というのは、検索サイトが Web ページの一部として提示するスニペットの持つテキスト文字数、120 字前後より明らかに多い。このことから、本論文が提案するブロック単位での提示が有用であることが期待される。

#### 最適なテキスト量のページ種別依存

図 7 より、Blog.1、Wiki.1、Wiki.2 のページにおいて、他のページに比べて最高評価値が低いことが分かる。また、集計した実験結果より、Blog.2、News.1、News.2 では、全員が「どちらかと言えば適切である」以上の評価を付けているが、Blog.1、Wiki.1、Wiki.2 では評価が分かれていることが分かった。

評価が分かれず良い評価を得られたページは、公式ブログとニュース記事のページである。これらのページでは文章や箇条書きのリストが段落ごとに成形され記述され、ある程度のまとまりごとにレイアウト構造が分かれている。

一方評価が分かれたページは、個人ブログと Wikipedia の 2 つの記事である。個人ブログである Blog.1 ではニュース記事を引用してページを作成しているので、段落ごとに成形されて記述されている。しかし記事を部分的に引用しているため、ニュース記事のページに比べブロックのサイズが大きすぎるとは言えないものになっている。このためブロックが多少大きくても意味のまとまりを重視する被験者と、意味のまとまりより提示に適したサイズを優先する傾向のある被験者とで評価が分かれた。また記事引用部分と広告部分が隣接してレイアウトされているため、階層構造も近くなり、ひとまとまりにされるという問題もあった。

今回実験に用いた Wikipedia のページには、箇条書きが多く含まれていた。長く連なった箇条書きは提示に適さないため分

割が必要だが、本研究ではテキスト量と階層構造にだけ着目して分割を行ったため、長く連なった箇条書きを意味のある部分で分割することが困難であった。このため被験者による評価が伸びなかったと考えられる。

また、被験者自身による分割が難しかったページを問うたアンケートでも、長い箇条書きの分割が難しいという理由で、6人中3人がWikipediaの記事ページを挙げている。このことから箇条書きの分割は、人の手による分割でも難しいことが分かった。一方でWikipedia記事の分割が簡単だったと答えた被験者も2人おり、評価が分かれた。難しかったと答えた被験者は、分割が難しくても提示に適したサイズで箇条書き部分も分割していたのに対し、簡単だったと答えた被験者は、ブロックのサイズが多少大きくても、意味の区切りで分割する傾向が見られた。このことから、箇条書きのブロックに関してはテキスト量よりも、意味のまとまりを優先させるべきであることが考えられる。

このように、いくつかの結果について評価が分かれており、これらのページほど評価値があまり高くならなかった。しかし、評価の分かれるページにおいても最高評価値は4.0を超えており、ある程度適切な分割が行えたとの結果を得られる。評価の別れが少なかったページにおいてはさらに高い結果が得られるなど、テキスト文字数300~500字で分割されたブロックが、ページ種別に依存することなく部分提示に十分適切なサイズとして用いることができると考えられる。

## 6. おわりに

本論文では、多くの情報が含まれるWebページの一部分を、ページのレイアウト情報を保持しつつ、あらかじめ指定した分量の情報を持ったブロックを単位に分割する手法を提案・実装した。本手法により分割されたブロックを単位として利用者に提示することで、ページ内から効率的に情報を発見できると考える。

ブロック分割に関する利用者実験を行い、本論文で提案した手法を用い300~500字のテキスト量で分割することで、提示に適切な大きさのブロックに分割できることが分かった。これはページ種別に依存することなく適切なサイズが得られるものとするが、今後更なる調査と検討が必要である。

今後は、Webページ分割の改良が必要である。改良点として、例えば結合相手がおらずブロックが小さいままである場合や、Webページ中に存在する画像や動画ファイルの扱いなどが挙げられ、それに伴う分割に用いるタグ選定の見直しも必要であると考えられる。改良後は、本手法により生成されたブロックを実際に提示するシステムを構築・評価することで、ブロックを単位とした利用者への提示が、情報発見において有用であることを示したい。

## 文 献

- [1] 総務省情報通信政策研究所. インターネット検索エンジンの現状と市場規模等に関する調査研究. 2009.

- <http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2009/2009-I-14.pdf>
- [2] 総務省情報通信政策研究所. ブログの実態に関する調査研究. <http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2009/2009-02.pdf>
- [3] 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.
- [4] 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 (JSAI 2009), 2B3-1. 2009.
- [5] 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.
- [6] 山本浩司, 山田誠二. Webページの部分表示によるPDAへの対話的Web適応. 情報処理学会研究報告. Vol.2002, No.105, pp.21-24, 2002.
- [7] Yu Chen, Wei-Ying Ma, and Hong-Jiang Zhang. Detecting web page structure for adaptive viewing on small form factor devices. In WWW '03: Proceedings of the 12th international conference on World Wide Web, pp.225-233. 2003.
- [8] 韓浩, 徳田雄洋. Web部分情報抽出システムとその応用. 日本ソフトウェア科学会第23回大会, 4B-1, 2006.
- [9] 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.