リンク元コンテキストを用いた WEB 文書の最重要箇所の同定法

小谷 忠史 *　　林 直弘 **　　鍋島 英知 ***　　岩沼 宏治 ***

* (株) 日本コンピュータコンサルタント
** 山梨大学 大学院 医学工学総合教育部
*** 山梨大学 大学院 医学工学総合研究部

{nabesima, iwanuma}@iw.media.yamanashi.ac.jp

概 要

本稿では，リンク元コンテキストに基づきリンク先 Web ページの最重要箇所を同定する手法を提案する．適切なコンテキストを抽出するため，リンク元ページにおける繰り返し構造に着目し，テキストの修飾情報に基づいて繰り返しの基礎単位を抽出するアルゴリズムを示す．本アルゴリズムは，実用時間内で高速に繰り返し構造を同定することが可能である．評価実験の結果，ニュースサイト及びテキストを主体とするリンクにおいて本手法が有用であることを示す．

キーワード： ブラウジング支援，繰り返しパターン，コンテキスト情報，WWW，プロキシサーバー

1　はじめに

現在，インターネット上には膨大な情報が溢れており，目的の情報を効率良く発見することは極めて困難になっている．一般にユーザーは，Google 等の検索エンジンを利用して 候補 Web ページを絞り込み，それらを閲覧して情報を獲得する．その際，ユーザはハイパーリンク（以後，単に**リンク**と呼ぶ）を辿りながら，数多くの Web ページを閲覧することになる．よって必要な情報を効率よく獲得するためには，検索エンジンがユーザに適切な Web ページ群を提示するだけでなく，Web ページ中のどこに注目すれば良いのかを提示することも重要である．例えば “松井の打点トップに 1 差” というリンクを辿る場合，ユーザがリンク先ページにおいて閲覧したい情報は，その記事の詳細であり，他の記事や広告等には興味がない．このような閲覧の意図を考慮したページ重要箇所の同定は，携帯端末のような小さな表示画面を用いる場合，特に極めて本質的かつ実用上も重要な問題となる．

そこで本論文では，Web ページ上のリンクを辿る際に，リンク先 Web ページにおいて最も重要であろう内容が書かれている箇所を自動的に同定する手法を提案する．リンク先 Web ページにおいて，ユーザーにとって何が重要であるかを判定するためには，ユーザがリンクをクリックするに至った経緯（**コンテキスト**）を把握する必要がある．我々はリンクを辿る際のコンテキストを，そのリンク周辺の文字列から抽出することを試みる．具体的には，幾つかの HTML タグを解析し，ページ中の繰返し構造等を解析し，リンクを含む繰返し基本部分をコンテキストと考える．そしてリンク先のページの単語の出現頻度その他を解析し，コンテキストと最も関連度の高い箇所をユーザに提示する．本手法の特徴は，実用時間で繰返し構造と重要箇所を精度良く高速に同定するところにある．

我々は提案した重要箇所の同定法を基に，ブラウジング支援システムを開発している．支援システムは WWW プロキシサーバーとして提供され，ユーザは好みのブラウザにより透過的に Web ページにアクセスできる．リンクをクリックすると，プロキシサーバーはリンク先ページの重要箇所を高速に同定

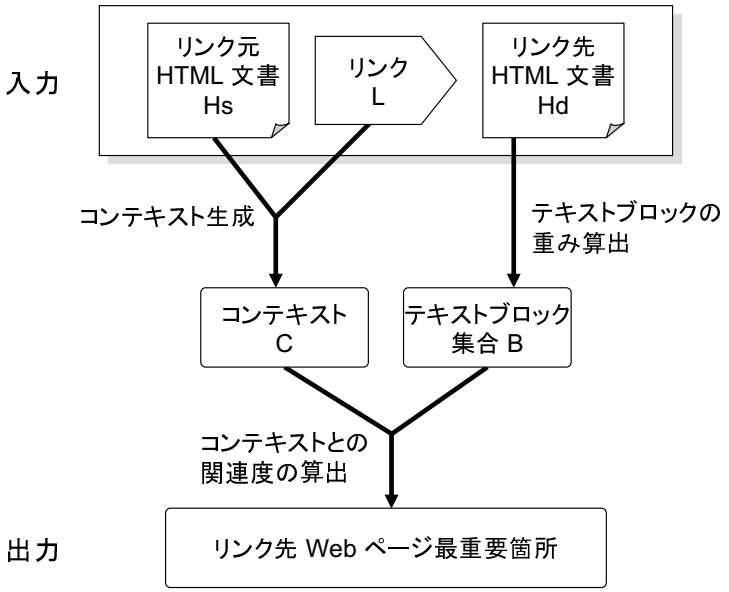

図 1: 最重要箇所同定までの処理の流れ

し，ブラウザの一番上に表示するように調整する．テキストを主体とする 4 個のサイトの計 80 リンクに対する評価実験の結果，約 70%以上のリンクについて正しく重要箇所を特定に成功した．

関連研究として，[南野 03] でも繰返しの単位を囲むタグが同じになることに着目し，多重の繰返し構造を再帰的に抽出する手法を提案している．また Web ラッパーの自動生成 [Hsu 97],[Kushmerick 97] や Web ページ差分検出 [Lim 01] でも，HTML タグを文書解析の手がかりとして使用している．一般にこれらの手法は多くの計算時間が必要であり，実時間での応答を要求するブラウジング支援には不向きである．[池田 03] は，部分文字列の出現頻度の差に基づき共通パタンを抽出する方法を提案している．この手法はパタンを抽出するために複数の Web ページを必要とするが，我々の手法はただ 1 つの Web ページから繰返し構造を抽出する．リンク解析は Web コミュニティ発見問題 [村田 02] やページのレーティングの基礎となる重要技術であるが，本研究で提案するページ中の重要箇所の同定法は，このリンク解析の改善にも寄与できる．リンクがページ内のどの部分を指しているのかを知ることで，より適切なページ間の関係を知ることが可能になると考えられる．最後に [増田 03] では，モバイル環境の表示の技術として， 文書要約の一種としての HTML 表データの認識などが研究されているが，本研究のようなリンク元コンテキストを考慮した重要部分の同定とその表示に関する研究は，これまでのところ筆者の知る限り，他に類似の研究は無いと思われる．

2 コンテキストを考慮した重要箇所同定法

HTML タグから次の HTML タグで囲まれているテキスト部分を**テキストブロック**と呼ぶ．特にアンカータグ <a>〜 に挟まれたテキストブロックを**アンカーテキスト** と呼ぶ．

本論文で提案する重要箇所同定プログラムの概要を図 1 に示す．プログラムは入力として，リンク元の HTML 文書 H_s とリンク先の HTML 文書 H_d，そして H_s と H_d を結ぶリンク L を受け取る．L は H_s 内のハイパーリンクである．まずプログラムは，L の周辺の適当な範囲の文字列をコンテキスト C として抽出する．次に C と H_d 中の各テキストブロックとの関連度を求め，最も関連度が高いテキストブロックを含む部分を最重要箇所として出力する．以下では詳細を述べる．

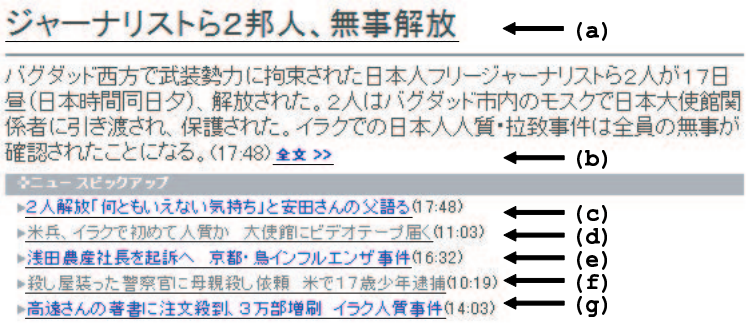

図 2: asahi.com トップページの一部

図 3: amazon.com の Web ページの一部

2.1 コンテキストの生成

本研究では，ユーザがリンクを辿るときの**コンテキスト**を，そのリンク周辺の文字列と考える．通常，Web ページは人間が理解しやすいようにレイアウトされる．例えば図 2 は一部に繰返し構造がある例である．(c)〜(g) はニュースのヘッドラインの繰返しであり，コンテキストはリンクのアンカーテキストそのものと考えられる．(a) と (b) は繰返し構造には含まれていない． 図 3 では，“表紙画像”，“タイトル”，“著者”，“価格” が順に繰り返されている．このような繰返しに含まれるリンクのコンテキストは，そのリンクを含む繰返しの単位とするのが自然である．繰返し構造を抽 出するために HTML タグに着目する．繰返し単位ではその修飾 HTML タグは同一になると考えられる．そこで，テキストの修飾タグの出現状況 が等しい最大のテキストブロック列を**繰返し単位**として定義する．

本論文では，適切なコンテキストを抽出するために，まず (1) 対象リンクが繰返し構造に含まれるか否かを判断し，もし含まれるならば，その繰返し単位をコンテキストとして抽出する．含まれない場合は，(2) 対象リンクの前後の一定範囲のテキストブロックをコンテキストとして抽出する．

繰返し単位を抽出するため，あるテキストブロック t を囲む HTML タグの集合を，そのテキスト

ブロックの**修飾値**と呼び，$Ad(t)$ と表す[1]．テキストブロックの列 $U = \langle u_1, \cdots, u_m \rangle$ の修飾値 $Ad(U)$ を $\langle Ad(u_1), \cdots, Ad(u_m) \rangle$ とする．またテキストブロック t_j よりも前にある k 個のテキストブロック列 $\langle t_{j-k}, \cdots, t_{j-2}, t_{j-1} \rangle$ を $H(t_j, k)$ と書く．同様に t_j よりも後ろにある k 個のテキストブロック列 $\langle t_{j+1}, t_{j+2}, \cdots, t_{j+k} \rangle$ を $T(t_j, k)$ と書く．

繰返し単位の抽出アルゴリズムを以下に示す．リンク元 Web ページのテキストブロックの集合を $T = \{t_1, \cdots, t_n\}$ とする．ユーザがクリックしたアンカーテキストを $t_{anc} \in T$ とする．

1. $A(\subseteq T)$ を，t_{anc} を含む最も内側の `<table>` と `</table>` に囲まれたアンカーテキストの集合とする．t_{anc} と等しい修飾値を持つ A 中のアンカーテキストの集合 $U = \{t_k \in A \mid Ad(t_k) = Ad(t_{anc})\}$ を求める．

2. 次式を満たす最大の値 a と b を求める．

$$\begin{aligned} \forall u_j \in U \ (Ad(H(u_j, a)) &= Ad(H(t_{anc}, a))). \\ \forall u_j \in U \ (Ad(T(u_j, b)) &= Ad(T(t_{anc}, b))). \end{aligned}$$

3. 繰返し単位として，以下のテキストブロック列

$$\langle t_{anc-a}, \cdots, t_{anc-1}, t_{anc}, t_{anc+1}, \cdots, t_{anc+b} \rangle$$

を出力する．

(1) では，考慮対象のアンカーテキストを，最も内側の `<table>` タグに囲まれた範囲に限定している．通常，繰返しの単位は連続して出現する．また近年の多くの Web ページでは， ページレイアウトのため に `<table>` タグが利用されているため，我々は同一の `<table>` タグに囲まれた領域のみを検査することとした．

もし $U = \{t_{anc}\}$ の場合，すなわちアンカーテキストが唯一つしか存在しない場合，または (2) において $a = b = 0$ となる場合，繰返し構造は存在しないと判断し，リンク周辺の一定範囲のテキストブロックをコンテキストとして抽出する．具体的には，t_{anc} から近い順に 3 つのテキストブロックを結合したものをコンテキストとする ．これは，5 つのニュースサイトの計 100 個のリンクに対する予備的実験の結果，t_{anc} から 3 つのテキストブロックをコンテキストとすると良好な結果が得られることが分かったためである [小谷 03]．

2.2 リンク先ページの解析

リンク先 Web ページにおいてコンテキストと最も関連する部分を特定する．まず初めに，全てのテキストブロックを $TF \cdot IDF$ により重み付けを行う．

ある Web ページ中のテキストブロック d における単語 t の出現数を $tf(d,t)$ で表し，単語 t が出現するテキストブロック数を $df(t)$ で表す．**テキストブロック d における単語 t の重み** $TF \cdot IDF(d,t)$ を次式で定義する．

$$TF \cdot IDF(d,t) = tf(d,t) \times \log \frac{N}{df(t)}$$

テキストブロック d の重み $W(d)$ を，d 中に出現する名詞[2]の重みの総和として定義する．

$$W(d) = \sum_{t \in d} TF \cdot IDF(d,t)$$

[1]本論文では，テキスト修飾タグのうち `<font>`, `<a>`, `<b>`, `<h1>`〜`<h6>` のみを利用した．予備調査の結果，これらのタグの使用頻度と繰返し区切り記号としての使用が多かったためである．また修飾タグの順序の相違は無視する．すなわち，`<a><b>foo</b></a>` も `<b><a>foo</a></b>` も同じ修飾値を持つと考える．

[2]名詞の抽出には日本語形態素解析ソフト茶筌 [松本 03] を利用した．

(a) リンク元ページ
（矢印はユーザがクリックしたリンク）

(b) 通常のリンク先ページ

(c) ブラウジング支援システム
によるリンク先ページ

図 4: ブラウジング支援システムの動作例

$W(d)$ は，テキストブロックの長さに比例して大きな値となる．一般的に，大きなテキストブロックほどそのページの主要なコンテンツになる傾向があるため，このような尺度を導入している．次に**コンテキスト C とテキストブロック d との関連度 $Rel(C,d)$** を以下で定める．

$$Rel(C,d) = W(d) \times Noun(C,d)$$

ここで $Noun(C,d)$ は，コンテキスト C に含まれる名詞がテキストブロック d に出現する頻度の総数を表す．上式は重要な単語を多く含み，かつコンテキストに含まれる名詞を多く含むテキストブロックほど，コンテキストと密接な関係にあるということを表現している．

近年の Web ページでは，レイアウトのために HTML のテーブルタグ（`<TABLE>`, `<TR>`, `<TD>` など）を多用している．また一般に，最も大きな面積を占めるセル（**セル**とは `<TD>` タグで囲まれたもの）が，その Web ページの重要箇所であることが多い．例えば，NewsBlaster [McKeown 02] はニュースサイトから新聞記事を自動収集し，要約を作成するシステムであるが，このシステムは新聞記事を自動収集する際，512 文字以上を含む Web ページ中の最も大きなセルを新聞記事と考えて抽出している．そこで本手法では，最も関連度の高いテキストブロック d' を含む最も内側のセルを，その Web ページの**最重要箇所**として出力する．もし d' がセルに含まれていない場合は Web ページ全体を出力する．

3　コンテキストを考慮するブラウジング支援システム

我々は，本稿で提案した最重要箇所同定法に基づくブラウジング支援システムを開発した．支援システムは WWW プロキシサーバーとして提供され，ユーザは普段から使い慣れたブラウザにより透過的に Web ページにアクセスできる．システムは Java 言語で実装されているため可搬性が高く，様々なプラットフォームで利用可能である．支援システムの動作例を図 4 に示す．図 4 (a) の矢印で示したリンクをクリックすると，通常 (b) の画面に移動するが，支援システムを通して利用した場合，リンク先ページの最重要箇所がブラウザの一番上に表示される（図 4 (c)）[3]．この重要箇所の同定処理は実用時間内に高速に処理され，数秒で処理が完了する．これによりブラウザのスクロールバーを操作する手間が軽減され，目的の情報にたどり着くまでのブラウジング作業が円滑になる．特に携帯端末を使って PC 用に設計された Web ページを閲覧する際には，小さな画面全体に広告や関係のないコンテンツが表示されることを避けることができ，ブラウジングの快適性を大きく向上させることができる．

[3]本システムでは，リンク先ページの重要箇所をブラウザの一番上に表示させるため，`<a>` タグの `name` 属性を利用している．

表 1: 繰返し単位の抽出実験結果

サイト名	成功	一部	失敗
Yahoo!ニュース	25	0	0
Asahi.com	25	0	0
アマゾン	21	4	0
楽天	19	6	0

4 評価実験

まず最初に，繰返し単位抽出アルゴリズムの評価実験を 4 個の Web サイトの計 100 リンクに対して行った（表 1）．表中の “成功”，“ 一部”，“失敗” は，それぞれ「繰返しの単位を正しく抽出できた」,「一部抽出できた」,「抽出できなかった」ことを示す．多くのリンクにおいて本手法が繰返しの単位を正しく抽出できていることが分かる．

次に 10 個の Web サイトの計 200 個のリンクについて最重要箇所同定法の評価実験を行った（表 2）．評価は，人手によって，リンク元のコンテキストを考慮した上で，リンク先の最重要箇所を正しく同定できているかどうかを判断した．表中の “成功”，“一部”，“失敗” は，それぞれ「最重要箇所を正しく同定できた」,「一部同定できた」,「同定できなかった」リンクの数を表す.「一部同定」の意味は，本来の最重要箇所が複数のセルにまたがっていた場合にその一部を自動同定できた，というものである．括弧内の数字は，本手法において繰返し構造を発見できなかったリンクの数を表している．

ニュースサイトにおいては非常に良い結果が得られた．その理由は，リンク先のページでは，記事本文の長さがその他のテキストに比べて長く，記事本文とコンテキストとの関連度が相対的に高くなったためである．ショッピングサイトでは最重要箇所が複数セルにまたがっている場合が多かったため，主に一部同定という結果となっている．最重要箇所の同定に失敗したリンクの多くは，アンカーテキストが「和書」や「工学部」などの場合で，リンク先のページ全体を指し示すようなリンクの場合や，リンク先が画像や Flash などの非テキストメディアで表現されているリンクである．そこでアンカーテキストが文章 になっているリンクで，かつ，リンク先がテキスト主体となっているようなリ ンクに対し実験を行った（表 3）．約 70%のリンクにおいて正 しく最重要箇所が同定でき，一部同定も含めると 95%以上の正当率を得ている．これより，特にテキストを主体とするページ間のリンクにおいて，本手法が極めて有効であることが分かる．

5 まとめ

本稿では，リンク元コンテキストに基づきリンク先 Web ページの最重要箇所を同定する手法を提案した．適切なコンテキストを抽出するため，リンク元ページにおける繰返し構造に着目し，テキストの修飾タグ情報に基づいて繰返しの基礎単位を抽出するアルゴリズムを提案した．本アルゴリズムは，ブラウジング実用時間内で高速に繰返し構造を同定することが可能である．評価実験の結果，ニュースサイト及びテキストを主体とするリンクにおいて本手法が有用であることが示された．

今後の課題の 1 つに，リンク先が画像や Flash などの非テキストメディアの場合に重要箇所を正しく同定する方法の検討が挙げられる．実用時間内に画像認識を行なうことは困難であるので，画像サイズや画像のリンク先等を手がかりとする方法が考えられる．また本手法では，最重要箇所の単位をセルとしているが，Web サイトによっては重要箇所が複数のセルにまたがる場合や，逆にセルの一部分のみである場合もある．より正確に重要箇所を特定することも今後の課題の 1 つである．

表 2: 最重要箇所抽出評価実験結果

カテゴリ	サイト名	成功	一部	失敗
ニュース	Yahoo!ニュース	18	2	0
	Asahi.com	17	3	0
	MSN	16	4	0
買物	アマゾン	0	8(3)	12(6)
	楽天	0	15(6)	5(2)
	goo ショッピング	0	13(5)	7(2)
企業	トヨタ自動車	2	4(1)	14(6)
	JAL	4	6	10
その他	山梨大学	6	3(1)	11(5)
	甲府市	4	4	12(5)
	計	67	62(16)	71(26)

表 3: テキスト主体のリンクにおける実験結果

カテゴリ	サイト	成功	一部	失敗
企業	トヨタ自動車	14	5	1
	JAL	12	6	2
その他	山梨大学	16(3)	4(2)	0
	甲府市	14	6(2)	0
	計	56(3)	21(4)	3

目標の情報に辿り着くまでに複数のリンクを辿ることはしばしばある．本手法では 1 つ前の Web ページにおけるリンクのコンテキストのみを考慮していたが，多重にリンクを辿る際のコンテキストを検討することで，より正確に重要箇所が特定できる可能性がある．このリンクを多重に辿る場合のコンテキストの生成も興味深い重要な課題である．

参考文献

[Hsu 97] J.Y.Hsu and W.Yih: Template-Based Information Mining from HTML Documents, *Proc. of AAAI-97*, pp.256-262 (1997).

[池田 03] 池田大輔, 山田泰寛, 廣川 佐千男：部分文字列増幅法による共通パタン発見アルゴリズム, 情報処理学会研究報告 2003-MPS-47 (2003)

[小谷 03] 小谷忠史, 岩沼宏治, 鍋島英知: リンク元コンテキストを考慮するハイパーリンク重要箇所同定法, 情報処理学会研究報告 2003-DD-39 (2004)

[Kushmerick 97] N.Kushmerick, D.S.Weld and B.Doorenbos: Wrapper Induction for Information Extraction, *Proc. of IJCAI-97*, pp.729-735 (1997)

[Lim 01] S-J.Lim, Y-K.Ng: An Automated Change-Detection Algorithm for HTML Documents Based on Semantic Hierarchies, *Proc. of ICDE 2001*, pp.303-312 (2001)

[McKeown 02] McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., Sable, C.. Schiffman, B., and Sigelman, S.: Tracking and Summarizing News on a Daily Basis with Columbia ' s Newsblaster, *Proc. of HLT-02* (2002)

[増田 03] 増田英孝，塚本修一，安富大輔，中川裕志：HTML の表示形式データの構造認識と携帯端末への応用, 情報処理学会論文誌：データベース， TOD.Vol.19, pp.23-32 (2003).

[村田 02] 村田剛志, ハイパーリンクのグラフ構造に基づく Web コミュニティの洗練人工知能学会誌, Vol.17, No.3, pp.322-329 (2002).

[南野 03] 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 2003-NL-154 (2003)

[松本 03] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: 形態素解析システム『茶筌』 version 2.3.3 使用説明書, http://chasen.aist-nara.ac.jp/, (2003)