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

小谷忠史†, 岩沼宏治††, 鍋島英知††

概要: 現在，インターネット上には膨大な量の情報が溢れており，ユーザが必要な情報に効率的にアクセスすることを支援する技術が求められている．ユーザが閲覧中の Web ページの重要箇所を自動的に提示できれば，要点の迅速な把握が可能となる．本論文では，ハイパーリンクを辿る際のコンテキストを考慮し，リンク先ページを解析することで重要箇所の同定する手法を提案する．評価実験の結果，ニュースサイトの記事抽出において良好な結果が得られた．

Identification of Important Region in Web Pages linked with HyperLink by Considering Context Information

Tadashi Kotani†, Koji Iwanuma††, Hidetomo Nabeshima††

Abstract: The Internet contains immense quantity information, and we need a technology which supports efficiently access to the desired information. In this paper, we propose how to identify an important region in Web pages linked with hyperlink by considering context information. Experimental results show that our proposed approach is effective for news sites.

1 はじめに

現在，インターネット上には膨大な量の情報が溢れており，ユーザが自身の目的にあった価値ある情報を効率良く見つけ出すことはますます困難になってきている．ユーザは必要な情報を入手するために，通常，Google や Yahoo! などの検索エンジンを利用して情報が含まれている Web ページを絞り込み，それらを閲覧しながら情報を獲得していく．また時事問題等の新鮮な情報は，ニュースサイト等を閲覧することにより獲得している．多くの場合，目的の情報にたどり着くまでにユーザは数多くの Web ページを閲覧することになる．効率の良い情報獲得のためには，検索エンジンが適切な Web ページをユーザに提示するだけでなく，ユーザが閲覧中の Web ページの重要箇所を特定することも重要である．

そこで我々は，Web ページ上のハイパーリンクを辿る際に，リンク元のコンテキストを考慮することにより，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部分箇所を自動的に同定する手法を提案する．

例えば，ニュースサイトにおいて “松井の打点，トップに１差” と記述されているハイパーリンクを辿る場合，リンク先のページにおいてユーザが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．また，ハイパーリンクが “続き” や “記事全文” と記述されていた場合には，ユーザはそのリンクの前後の文脈（**コンテキスト**）からリンク先のページにある記事を推定し，興味があれば，その記事の詳細のみを閲覧したい．

人間はコンテキストを容易に理解することができるが，機械にコンテキストを理解させるためには高度な自然言語処理が必要であり困難である．そこで我々は，ハイパーリンクを辿る際のコンテキストを，

†山梨大学大学院工学研究科コンピュータ・メディア工学専攻
†Department of Computer Science and Media Engineering
††山梨大学大学院医学工学総合研究部
††Interdisciplinary Graduate School of Medicine and Engineering, Yamanashi University

表 1: 実験対象の Web サイト

サイト名	リンク数	URL
asahi.com	150	http://www.asahi.com/
goo news	107	http://news.goo.ne.jp/

リンクの前後の文字列として単純に定義する．そして，そのコンテキストと最も類似度の高いリンク先ページの部分箇所をユーザに提示する．システムを実装し，評価実験の結果，多くのニュースサイトにおいて良好な結果を得ることができた．

本論文の構成は以下の通りである．まず 2 章において，コンテキストの定義について述べ，コンテキストとして考慮すべきテキストの範囲に関する予備実験結果を示す．3 章ではコンテキストを考慮した重要箇所抽出方法について述べ，4 章においてその評価実験の結果を示す．5 章で関連研究を紹介し，6 章で本研究をまとめ，今後の課題を示す．

2 ハイパーリンクのコンテキスト

ハイパーリンクを辿る場合のコンテキストを定義するため，補助的な定義を行う．HTML 文書を任意の HTML タグで区切り，タグ以外の各文字列を**テキストブロック**と呼ぶ．アンカータグ <a> 〜 内のテキストブロックのことを特に**アンカーテキスト**と呼ぶ．あるアンカーテキスト A とテキストブロック T との**距離** $D(A,T)$ を，A から T までの間に存在する HTML タグの数として定義する．

HTML 文書内のあるアンカーテキストを A とし，その前後にあるテキストブロック列を $T_1, T_2, \ldots, T_{i-1}, A, T_{i+1}, \ldots, T_n$ とする．コンテキストとして考慮するテキストブロックの範囲を k とする．$k \geq 0$ である．**アンカーテキスト A の範囲 k のコンテキスト** $C(A,k)$ を次式で定義する：

$$C(A,k) = A \cup \{T_j \mid D(A,T_j) \leq k\}$$

すなわちコンテキスト $C(A,k)$ とは，アンカーテキスト A と，A との距離が k 以下のテキストブロックの集合である．

我々は，コンテキストとして適切な範囲 k を知るため，次のような予備実験を行った．表 1 に示す 2 つのサイトのトップページは，合計 257 個のハイパーリンク（アンカーテキスト）を含んでいる．各アンカーテキスト A_i に対し，コンテキスト $C(A_i,0), \ldots, C(A_i,5)$ を生成し，これらから，コンテキストとしてふさわしい $C(A_i,t)$ を選んだ．ここで，コンテキストとしてふさわしいかどうかは，実際に人間がリンクを辿る際にコンテキストと考慮しているテキストブロックを判断基準とした．その結果を表 2 に示す．表中の各要素は，コンテキストとしてふさわしいテキストブロックの数を表している．例えば，asahi.com のコンテキストの範囲が 0 のとき 55 個というのは，コンテキストとして考慮するテキストブロックがアンカーテキストのみと判断したリンクが 55 個あったことを表す．また，その他とは，コンテキストの範囲が 6 以上の場合，またはコンテキストとしてふさわしいテキストブロックが無い場合である．

表 2: コンテキストの評価

	コンテキストの範囲 k						
	0	1	2	3	4	5	その他
asahi.com	55	46	24	9	0	0	16
goo news	66	17	7	5	0	0	12
合計	121	63	31	14	0	0	28

表 2 の結果は，アンカーテキストの前後の 3 ブロックをコンテキストとして考慮すれば，asahi.com の場合は 150 個のリンクのうち 134 個 (89.3%) について適切なコンテキストを生成することが可能であり，goo new の場合は 107 個のリンクのうち 95 個 (88.7%) について適切なリンクを生成することが可能であることを示している．

3 コンテキストを考慮した重要箇所同定法

我々の提案する重要箇所同定法の概要を図 1 に示す．入力として，リンク元ページとリンク先ページ，それらのページを結ぶリンクとを与える．次に，そのリンクからコンテキストを生成する．我々の手法では，先の予備実験の結果から，範囲 3 のコンテキ

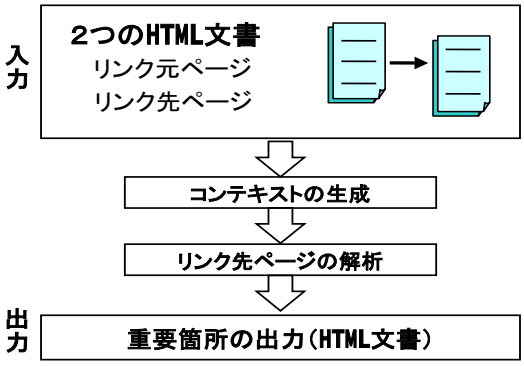

図 1: 本手法の概要

ストを生成する．その後，リンク先ページを解析し，コンテキストと最も関連性のある部分箇所を特定し，それをリンク先ページの重要箇所として出力する．本節では，リンク先ページの解析と，重要箇所の特定方法について説明する．

3.1 リンク先ページの解析

リンク先ページにおける重要箇所を特定するため，まずリンク先ページのすべてのテキストブロックについてその重要度を算出する．テキストブロックの重要度とは，そのテキストブロックに出現する単語の重みの総和として表現される．まず単語の重みについて定義し，その後テキストブロックの重要度を定義する．

単語の重み付け

単語の重み付けには，TF·IDF 法を用いる．リンク先ページのテキストブロックの総数を n とし，テキストブロック d における単語 t の出現数を $TF(d,t)$，単語 t が出現するテキストブロック数を $df(t)$ としたとき，**テキストブロック d における単語 t の重み** $TF \cdot IDF(d,t)$ は次式により定義される．

$$TF \cdot IDF(d,t) = TF(d,t) \times \log \frac{N}{df(t)} \qquad (1)$$

ここで，重み付けされる単語は，各テキストブロックから抽出した名詞のみとする*．

*日本語形態素解析ソフトとして茶筌 [4] を利用した．

テキストブロックの重要度

テキストブロックの重要度は，テキストブロックに出現する単語の重みの総和として定義する．テキストブロックに含まれる名詞の数を m とし，テキストブロック d における i 番目の名詞を t_i とする．**テキストブロック d の重要度** $S(d)$ は次式で与えられる．

$$S(d) = \sum_{i=1}^{m} TF \cdot IDF(d,t_i) \qquad (2)$$

テキストブロックが長くなればなるほど，そのテキストブロックは名詞を多く含むので重要度は大きな値となる．一般に，テキストブロックに重要な単語が含まれ，かつ，テキストブロックが長いものほど重要度が高くなるといえる．

3.2 重要箇所の出力

前節で定義したテキストブロックの重要度とリンク元のコンテキストとの関連度を調べることにより，リンク先 Web ページにおける重要箇所を特定する．

コンテキスト C とテキストブロック d との**関連度** $R(C,d)$ を次式で定義する：

$$R(C,d) = N(C,d) \times S(d) \qquad (3)$$

ここで $N(C,d)$ は，コンテキスト C に含まれるすべての名詞がテキストブロック d に出現する総数を表す．我々は，重要度の高いテキストブロックであり，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考えた．

リンク先 Web ページのすべてのテキストブロックについて，コンテキストとの関連度を求め，最も高い関連度をもつテキストブロック d' を決定する．

ところで，近年多くの Web ページでは，レイアウトのために HTML のテーブルタグ<TABLE>，<THEAD>，<TBODY>，<TFOOT>，<TR>，<TH>，<TD> を利用している．さらに，多くの場合において，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とが多い．例えば NewsBlaster は，ニュースサイトから新聞記事を収集し，要約を自動作成する要約システムであるが，このシステムは新聞記事を自動収集する際，Web ページの最も

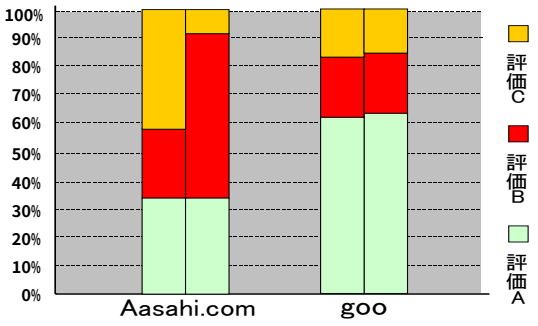

図 2: コンテキストの有効性

図 3: 入力 Web ページ：Yahoo! の例

図 4: 出力結果：Yahoo! の例

大きなセルの中で 512 文字を超えるものを新聞記事と考えて抽出している [3].

そこで本研究では，d' を含むセルをその Web ページの重要箇所として出力する．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．逆に，セルが入れ子になっている場合は，d' を囲む <TD> タグのうち，d' に最も近いセルを出力する．

我々は重要箇所の抽出のためにコンテキストとテキストブロックとの “関連度” を用いたが，テキスト処理においてよく利用されている尺度である “類似度” を利用する方法も考えることができる．一般にテキスト同士の類似度とは，各テキスト中に出現する単語の重みのベクトルの内積等で定義されることが多い．しかしリンクを辿る行為は，リンク元ページには無い情報をリンク先ページから得ることであるため，一概にコンテキストとリンク先ページのテキストブロックの類似度の値が大きければ良いとはいえない．そこで本手法では，コンテキストに含まれる名詞を多く含んでいるテキストブロックほど重要であるという尺度を採用した．

4 実験結果

表 1 の 2 つのサイトについて本手法を適用し，重要箇所の同定を行った．また比較のために，コンテキストを全く考慮しない単純な重要箇所の同定法についても実験を行った．その手法では，リンク先ページのテキストブロックの重要度が最も大きなテキストブロックを含むセルを重要箇所として出力する．重要箇所がすべて抽出できた場合を評価 A とし，全く抽出できなかった場合を評価 C とした．その他の場合評価 B である．本手法は重要箇所として 1 つのセルのみを出力するため，重要箇所が複数のセルに存在する場合，評価 B もしくは評価 C となる．

asahi.com，goo の各サイトのトップページから，それぞれランダムにリンクを 40 個選び，リンク先 Web ページの重要箇所を抽出させて評価を行った．実験結果を図 2 に示す．各サイトについて，左側はコンテキストを考慮しない単純な手法の結果であり，右側がコンテキストを考慮した手法の結果である．

評価 A については，両手法において大きな違いはない．これは，重要箇所が 1 つのセルに収まっているような Web ページに関しては，コンテキストを考慮せずテキストブロックの重要度のみに基づいた抽出手法であっても有効であることを示している．一方，評価 B についてはコンテキストを考慮した手法のほうが優れていることが分かる．重要箇所が複数のセルに分かれる場合，コンテキストを考慮することが重要であることがわかる．

また，本手法を 5 つのニュースサイトに適用した結果を表 3 に示す．それぞれのサイトにおいて 20 個のリンクについて実験を行った．ただし，それら

表 3: ニュースサイトにおける評価

サイト名	リンク数	URL	評価 A	評価 B	評価 C
asahi.com	20	http://www.asahi.com/	20	0	0
YAHOO! JAPAN	20	http://news.goo.ne.jp/	20	0	0
NIKKEI NET	20	http://nikkei.co.jp/	20	0	0
Excite news	20	http://excite.co.jp/News/	20	0	0
読売新聞社 sports	20	http://www.yomiuri.co.jp/sports/	20	0	0

図 5: 入力 Web ページ：nikkei の例

図 6: 出力結果：nikkei の例

のリンクは全てニュースの見出しであり，リンク先ページはその記事の全文であるようなリンクを選んだ．我々の手法は，どのサイトにおいても正しく重要箇所を抽出できていることが分かる．

図 3〜6 は，実験結果の例である．図 3 の左側は YAHOO!JAPAN のトップページであり，そのページにあるハイパーリンク “109 市長が決定へ統一地方選の後半戦開票” を辿ったページが右側である．図 4 は，右側ページにおける重要箇所を抽出した結果である．また同様に NIKKEI.NET について，ハイパーリンク “衆院東京 6 区，民主・小宮山氏が初当選” について実験を行った結果が図 5, 6 である．

5 関連研究

本研究では，コンテキストを，アンカーテキストと，そのアンカーテキストからの指定の距離内にあるテキストブロックの集合として定義した．そしてそのコンテキストと最も関連度の高いリンク先のテキストブロックを含むセルを重要箇所として出力する．

アンカーテキストの前後の文字列を利用して，リンク先を特徴付けている研究として Eric らの研究がある [1]．彼らは，Web ページを自動分類する際に，カテゴリーの名前付けにリンク元 Web ページのアンカーテキストの前後の文字列（アンカーテキストを含めて 25 文字；**拡張アンカーテキスト** (Extended anchor text) と呼ぶ）を用いている．

またリンク先を Web ページを特徴付ける研究として，荒木らは，アンカーテキストの前後の文字列ではなく，Web ページのリンク集からリンクに関連のある見出し語を抽出し，リンク先ページの特徴として情報検索に応用している [2]．

6 まとめ

本論文では，ハイパーリンク周辺のテキストをコンテキストとして考慮し重要箇所を特定する手法を提案した．我々の手法は，特にニュースサイトにおいて有効であった．今後の課題として以下の点が挙げられる．

(1) **コンテキストの改善** : 本研究で定義したコンテキストは，距離が近いものから順に加えていくため，全く関係のないテキストがコンテキストと加え

られる場合がある．従って，コンテキストに加えるテキストブロックを制限することで，重要箇所の抽出結果を改善できると考えられる．

(2) **テキストブロックの重要度の改善**：本研究では，リンク先ページの各テキストブロックの名詞について，TF·IDF 法により重み付けをしている．しかし，全ての名詞に重み付けする必要はない．例えば，ニュースサイトでは,「ニュース」「新聞」等の単語はあまり必要ではない．一般に文書は，文字列が長くなるほど名詞を多く含むので，今回の手法では各テキストブロックの重要度と長さは比例することになる．そのため，不要語となる単語を考慮する必要がある．

(3) **セルによる重要箇所抽出法の改善**：本研究では，重要箇所を抽出する際，リンク先ページから重要度の最も高いテキストブロックを発見し，そのテキストブロックを含むセルを抽出している．しかし，実験で用いたニュースサイトのように，テーブルタグによるレイアウトを使っている Web ページでは本手法は有効であるが，そうでない Web ページでは別の抽出法を検討する必要がある．例えば，関連度の高いテキストブロックをある閾値に基づいて抜き出すことも考えられる．

参考文献

[1] Eric J. Glover, Kostas Tsiousiouliklis, Steve Lawrence, David M. Pennock, Gary W. Flake : “Using Web Structure for Classiying and Describing Web Pages”, Proc. of Eleventh International World Wide Web Conference, pp. 567-569 (2002)

[2] 荒木 良，中島 伸介，角谷 和俊，田中 克巳 : “Web ページのアスペクトに基づくクラスタリングとその応用” 2002-DBS-128,pp289-296, 情報処理学会報告書，2002.

[3] “ Newsblaster”, `http://www.cs.columbia.edu/nlp/newsblaster/`

[4] 形態素解析ソフト「茶筅」: `http://chasen.aist-nara.ac.jp/`