報知的要約を用いたリンク先内容表示による Web ページ理解支援システム

Web-Text Comprehension Support System by Displaying Informative Summaries of Link-Connected Web pages

中村浩介 [1*]　砂山渡 [1]

Kosuke NAKAMURA[1] and Wataru SUNAYAMA[1]

[1] 広島市立大学大学院情報科学研究科

[1] Graduate School of Information Science, Hiroshima City University.

1　はじめに

我々が Web ページを読んで情報を理解する際，Web ページ内に貼られた関連する内容のリンク先を参照して理解を深めることが多い．その際，複数の Web ページを別々のウィンドウに表示すると，ディスプレイの大きさの制約によりそれらを同時に閲覧することができない．また，現在実用化されている「タブブラウザ」を含め，複数のウィンドウに表示された Web ページを切り替えながら参照することで複数の Web ページを読む形式も存在するが，情報理解という観点を考慮すると，閲覧中の Web ページとそのリンク先の Web ページの両方の関連を捉えつつ参照するため，画面の表示切替が頻出してしまい，情報参照の効率が下がる．

そこで，本研究では，閲覧中の Web ページと，参照可能性の高いリンク先の Web ページの内容を一画面内にレイアウトして表示することで切り替え不要な表示形式を用いる．また，リンク先の内容の表示方法として報知的要約を適用した文書を使用することで情報を圧縮し，複数の情報を同時に表示することで情報理解の効率を高めるインタフェースを提案する．尚，本研究では情報理解支援を「有効な情報を網羅的に集めること」と定義しその効果を検証していく．

2　関連研究

閲覧中の Web ページに添付された複数のリンク先を一度に閲覧することができる Web ブラウザとして，Elastic Windows [Kandogan,E. 1997] がある．

リンク先はユーザが指定し，閲覧中のページの右側に表示する．ユーザは閲覧中のページと指定したリンク先の表示される場所を確認しながら Web ページを閲覧できるが，表示する Web ページの数が増えるほど各 Web ページの表示領域が小さくなってしまうため，同時に多数の情報を深く参照することはできない．これに対し本研究では，表示するリンク先の内容として要約文を用いることで複数の Web ページの表示文章量を下げ，表示領域を確保する点で異なる．

WWW 情報探索において，リンク先の Web ページの要約を提示する手法として，先読み代理サーバを用いた WWW 情報探索支援 [新井 2002] がある．

このシステムはポップアップ形式で表示されたリンク先の簡易要約文により，ユーザは閲覧する前にどのリンク先が重要かを判断可能にする機能を備えている．簡易要約文の生成には処理時間が速い簡易要約器 Posum[2] を使用し，Web ページの閲覧にリアルタイムで表示を対応させるために，先読みを行っている．

簡易要約器による指示的要約では，リンク先の内容を読むかどうかの判断を補助することができるが，リンク先の原文を理解する上で必要な情報が省略されてしまう可能性がある．

本研究では，リンク先の情報量を維持することで理解支援を行うため，指示的要約ではなく原文の代替物として使用することを目的とした報知的要約を用いる．

3　Web ページ理解支援システム

3.1　システム概要

図 1 にシステムの概要を示す．以下，図 1 における各モジュール (HTML 処理部，要約生成部，インタフェース部) について説明する．

3.2　システムへの入力

本研究ではシステムに与える入力は，ユーザが閲覧対象とした Web ページのアドレスである．また，本研究で提案するシステムの適用対象となる Web ページの条件を以下に列挙する．

- 一つの項目について文章により述べている Web ページであること．

*連絡先：広島市立大学大学院 情報科学研究科 情報機械システム工学専攻
〒 731-3194 広島県広島市安佐南区大塚東 3 丁目 4 番 1 号
E-mail: k-nakamura@sys.im.hiroshima-cu.ac.jp

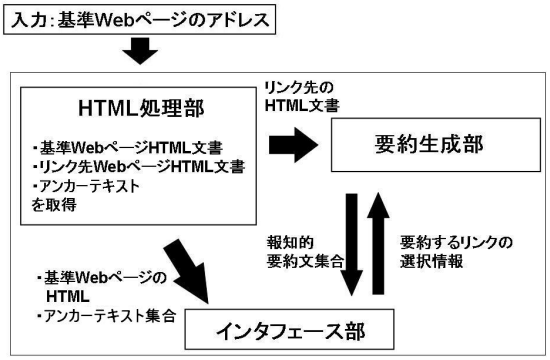

図 1: Web ページ理解支援システム

- Web ページ内に文字をアンカーとした他の Web ページへのリンクが貼られている Web ページであること．
- 閲覧中の Web ページのリンク先の Web ページは，文章を主体とした Web ページであること．

3.3 HTML 処理部

このモジュールは，ユーザが入力した Web ページのアドレス入力とし，その Web ページの HTML 文書のソースをダウンロードする．取得した HTML 文書を解析し，リンク先のアドレス，アンカーテキストを抽出する．その上で，リンク先のアドレスから HTML 文書を取得し，要約生成部に渡す．その後，アンカーテキスト集合，閲覧対象 Web ページの HTML 文書をインタフェース部へ渡す．

また，本研究では要約を適用し提示するリンク内容として，Web ページの本文と関係の強い本文中のリンクを対象とする．そこで，HTML 文書をブロックレベル要素のタグに基づいて複数の段落に分割し，各段落で

$$W_c = \frac{W_b}{W_a} \tag{1}$$

(W_b:アンカーテキストの文字数，W_a:段落内の全文字数）を求める．

式 (1) が一定値を超える段落に存在するリンクはリンク集として判断することができる．リンク集として判断される W_c の閾値は，パラメータとして経験則的に決定される．これは，リンク集としての機能を持つリンク集合の中には，リンク先の名前や一言の解説などを含むことがあり，リンク集である段落内での W_c は常に一定値を取るわけではないためである．リンク集として判断された本文との関係が弱いものとして，要約対象から除いた．

3.4 要約生成部

このモジュールは，HTML 処理部が取得したリンク先の HTML 文書を入力とし，報知的要約文を出力とする．

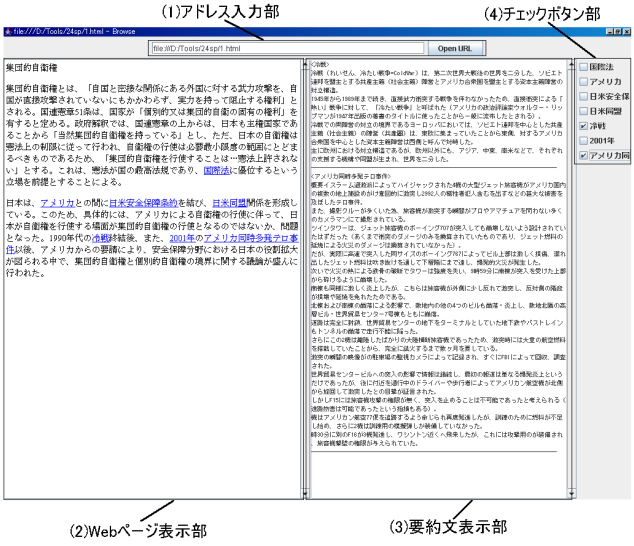

図 2: インタフェースの構成

報知的要約の一手法として相良らが提案したサブトピックを考慮した重要文抽出による報知的要約生成 [相良 2007] がある．この手法は，重要文抽出システムである展望台システム [砂山 2002] を拡張し，テキストのストーリーに基づく要約を作成するものである．この報知的要約は原文に対する要約の割合が 30 %のとき 80 %のあらすじ再現率を達成している．

本研究ではリンク先の内容にも話の流れを持たせるため，この要約手法を適用した要約文をリンク先の表示内容に使用した．

3.5 インタフェース部

システムの出力を表示するインタフェースは図 2 に示す 4 つの構成要素から成る．

(1) アドレス入力部

HTML 処理部に与えるアドレスを入力する．

(2) Web ページ表示部

HTML 処理部が取得した基準 Web ページ表示する．

(3) 要約文表示部

要約生成部で出力されたテキストを表示する．要約文表示部では複数の要約文を同時に表示することができ，表示領域はスクロールバーにより制御される．

(4) チェックボタン部

HTML 処理部が取得した各アンカーテキストの先頭 5 字を名前に持つチェックボタンを生成する．チェックボタンの ON/OFF で，各リンク先の要約文の表示非表示を選択できる．

表 1: 実験に使用した Web ページの用語

社会一般	おたく用語	社会用語
公害病	ショタコン	ワンセグ
集団的自衛権	ツンデレ	暗黒物質
日本の経済と金融	ウィキペたん	ニート

4 システム評価実験

前章で述べた提案システムが，Web ページを効率よく理解する上で支援効果があるかどうかを調べる実験を行った．

4.1 実験内容

評価実験における手順を以下に示す．

1. 被験者数：情報系の大学生及び大学院生 18 人
2. 被験者には「インタフェースを使用して，基準となる Web ページを理解する上で有効なリンク先のテキストを，簡潔かつ網羅的に抽出してください」という課題を与え，実際に使用したパソコン上の指定したテキストファイルにカットペーストにより抽出してもらった．
3. 被験者には上記の課題を 3 回行ってもらい，3 回のリンク先の表示にはそれぞれ要約なし，指示的要約，報知的要約の 3 種類の文章を提示した．

実験には，ウィキペディア [Wikipedia] の用語解説ページを元に，文中に複数のリンクを含む 9 種類の用語を持つ Web ページを用いた．表 1 に用語の一覧を示す．

評価実験は本研究で提案したシステムを含む 3 つのインタフェースを比較することで行った．各インタフェースは図 2 と同様の左右の表示領域を持ち，基準となる Web ページが左のウィンドウに表示され，右側の表示内容は (A, B, C) で異なる．また，各リンク先の表示と非表示はチェックボタンにより制御される．

A. リンク先の全文

B. リンク先の文書の指示的要約 (平均要約率:13 %)

C. リンク先の文書の報知的要約文
(平均要約率:27 %)

また，アンケートを行い，システムの使いやすさ，表示された文章の量や内容など感想を自由記述で回答してもらった．

表 2: 各表示方法における各用語で平均した提示文数と獲得文数

提示方法	提示文数 A	獲得文数 B	A/B
要約なし	307.78	20.17	15.26
指示的要約	35.44	10.73	3.31
報知的要約	83.11	14.67	5.67

表 3: 各用語での引用文数平均と引用文数の要約なしの引用文数に対する割合

提示方法	引用文数	要約なしの引用文数に対する割合
要約なし	20.17	1.00
指示的要約	10.73	0.53
報知的要約	14.67	0.73

4.2 実験結果

本実験では，被験者により引用された文数と被験者が回答時に選択したリンク先の数に着目した．得られた結果として，リンク先の各表示方法において各用語の実験課題で提示されたリンク先の文数の合計を平均した値，各用語での実験において被験者が回答時に引用した文数の合計を平均した値，及びそれらの割合，すなわち 1 文獲得するために参照された文数を表 2 に示す．

表 3 には各表示方法で課題時に被験者が引用した文数の各用語での平均値，及びその値について要約なしの値を基準としてそれぞれの表示方法での値の要約なしでの値に対する割合を算出した値を示した．

また，各被験者ごとの各提示方法における，1 文獲得するために参照した文数を昇順にプロットした図を図 3 に示す．

図 4 には，各被験者が課題回答時において，引用時に使用したリンク先の数の提示されている全てのリンクの数に対する割合を昇順にプロットした図を示す．

4.3 考察

4.3.1 提示文数と獲得文数の関係

表 2，図 3 より，各用語における課題において，引用された文数は指示的要約，報知的要約，要約なしの順に大きくなっていることがわかる．これは提示文章量が多いほど有効な情報として判断される文の量も増えることを示している．しかし，表 2 の 1 文獲得時に必要とする文数を参照すると要約なしの提示では 15.26 文と要約提示に比べて非常に大きな値を示している．これは，多量の文章が提示されている中から有効な情報を判断して引用するという行為が被験者に大きな負担となることを示している．

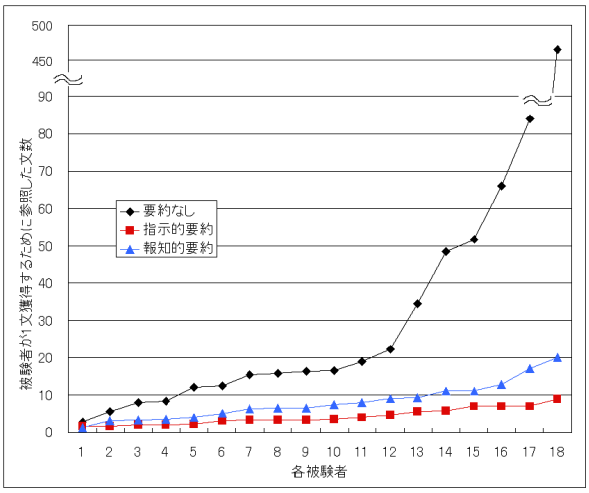

図 3: 各被験者が 1 文獲得するために参照した文数

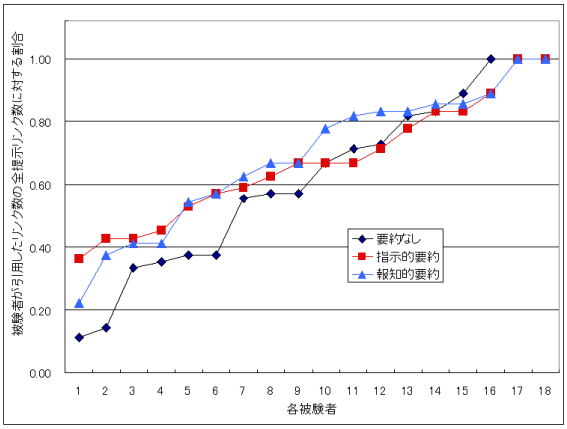

図 4: 各被験者が引用したリンク数の提示全リンク数に対する割合

また，表 3 の要約なし提示で引用された文数平均を基準とした各表示方法の引用文数平均の割合では，指示的要約は 5 割程度，報知的要約は 7 割程度の引用文数であることが示されている．指示的要約の提示では 13 %の要約率になるため，提示できる情報量には限りがある．そのため，要約なしに比べて有効と認識される文の数が半減してしまう．

報知的要約ではその要約率は 27 %であり，引用された文の数は 3 割程度の減少に抑えられているが，その情報量は要約なしの場合に対して大きく下がるものではないと言える．その理由として，被験者の引用方法が挙げられる．提示されている文章から有効な文を探す場合，被験者の多くは必要な部分を見つけるとそこからブロック単位で文を引用していた．これを考慮すると，自動要約手法では重要な単語に重みを付けることで価値の高い文を抽出することで，価値の高いブロックで構成された文章が出力されているため，情報量の減少は抑えられてるといえる．

その性質を持った上で，引用された文数の割合が 7 割を維持できていることから，報知的要約の提示が有効な情報を網羅的に集めるために効率的であることがわかる．

4.3.2 課題回答に使用したリンクの数

図 4 より，要約を提示した際に被験者は多くのリンク先から文を引用していることがわかる．これは 1 つ 1 つのリンク先がある程度まとめられていることによって，被験者が多くの情報に目が行き届くようになったことを示している．このことから，要約文の提示が閲覧中の Web ページのリンク先から有効な情報を網羅的に集めるという目的において効率的に働いていることがわかる．

4.3.3 アンケートの回答

被験者によるアンケートの記述中に,「リンク先を並べて表示できることは効果的である」,「報知的要約は詳しく調べる上で向いている」といった意見があり，提案システムの報知的要約による表示が有効な情報を網羅的に集める上で効果的といえる．

5 結論

システム評価実験の結果，及び考察より，文章量を 3 割程度に抑えつつ，被験者が閲覧中の Web ページを理解する上で有効と考えた文の量を 7 割に維持し，情報量の減少を抑えることを可能にした報知的要約文の提示が，リンク先を含めた Web ページの理解の支援が可能であることを示した．

参考文献

[Kandogan,E. 1997] Kandogan, E.and Shneiderman,B.: Elastic Windows: A Hierarchical Multi-Window World-Wide web Browser,Proceedings of the 10th annual ACM Symposium on User Interface Software and Technology(UIST'97),pp. 169-177(1997).

[新井 2002] 新井孝之，望月源，白井清昭，奥村学，先読み代理サーバを用いた WWW 情報探索支援，言語処理学会第 8 回年次大会 (2002).

[相良 2007] 相良直樹・砂山渡・谷内田正彦：サブトピックを考慮した重要文抽出による報知的要約生成，電子情報通信学会論文誌, Vol.J90-D, No.2, (2007).

[砂山 2002] 砂山渡・谷内田正彦：観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装，人工知能学会論文誌，Vol.17, No.1, pp.14 – 22, (2002).

[Wikipedia] フリー百科事典ウィキペディア (Wikipedia)URL(http://ja.wikipedia.org/wiki/).