先読み代理サーバを用いたWWW情報探索支援

新井 孝之† 望月 源† 白井 清昭† 奥村 学††
†北陸先端科学技術大学院大学 情報科学研究科
††東京工業大学 精密工学研究所

1 はじめに

近年，World Wide Web(以下 WWW と略す) の普及により，世界中に分散された文書に容易にアクセスすることが可能になった．しかし，その情報空間は膨大で，ユーザが知りたい情報をすぐに得られないことも多い．そのため，WWW における情報探索を支援する技術を開発することは重要な課題である．

通常，Web ページには多数のリンクが付けられており，ユーザはリンク先にジャンプしたり，また戻ってきたりして Web 探索を行う．しかし，この動作は非常に煩わしい．Greenberg によれば，Web ブラウジングにおいて，「戻る」ボタンを押すことは，すべてのナビゲーションイベントの30%以上に達する [1]．この「戻る」ボタンの使用を減らすこと，すなわち，リンク先に頻繁にジャンプするイベントを減らすことは，Web 探索の効率を上げるのに役立つと考える．

本研究では，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ことにより，どのリンク先が重要かをユーザが実際に閲覧する前に判断できるシステムを構築する．近年，WWW のように，電子化されたテキストが大量に利用可能になっていることから，テキスト自動要約は脚光を浴びてきている [2] [3]．本研究では，自動要約技術を WWW 情報探索支援に用いることを考える．また，ユーザが実際にページを参照する前に，事前にリンク先を取得する（先読みする）ことにより，ユーザがリンク先をマウスでクリックしてからリンク先のページが表示されるまでの時間を短縮する．

2 提案システム

2.1 概要

まず，提案する WWW 情報探索支援システムの大まかな処理の流れについて説明する．ユーザが Web ページを表示させると，そのページに存在するすべてのハイパーリンクについて，リンク先のページを先読みし，キャッシュに保存する．ユーザが実際にリンクをクリックしたら，キャッシュに保存されている情報を取り出すことによってリンク先のページを高速に表示できる．また，リンク先のページの要約を作成する．ユーザがマウスの上にマウスポインタを置いたとき，ツールチップとして要約を表示する．要約を表示することにより，ユーザはリンク先の情報が有用であるかどうかを判断することができる．これにより，ユーザがあまり関心のないページを表示させてから元のページに戻る操作を少なくすることが期待できる．ツールチップによる要約の表示例を図 1 に示す．

図 1: ツールチップ表示

要約を表示する手法としては様々なものがあるが，ツールチップによる要約表示には以下のような利点があると考えられる．

- 現在表示しているページ（参照元のページ）にリンク先の要約を埋め込んでユーザに提示する方法[1]では，要約作成に時間がかかる場合，ページの表示も遅れる．これに対し，ツールチップを用いれば，参照元のページと要約のページは別ページにできるので，要約作成に時間がかかっても，参照元のページは先に表示される．
- リンクの上にマウスポインタを置くという簡単な操作方法でリンク先の要約を表示するため，ユーザが利用しやすい．

[1]例えば，検索エンジンなどで，検索結果のリンクとリンク先の要約を同時に提示する場合は，リンク先の要約を参照元ページに埋め込んでいるとみなせる．

- ツールチップによって要約を表示しない限り，参照元のページのレイアウトは変更されない．

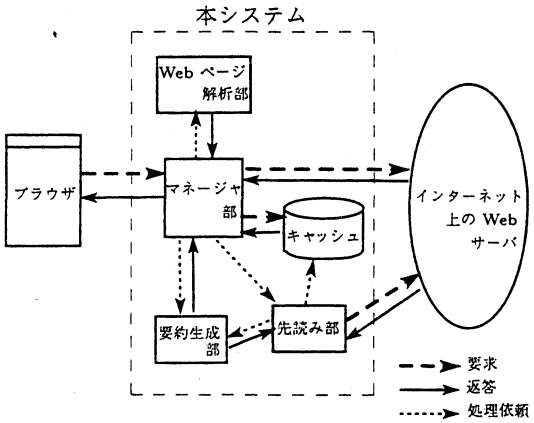

図 2: システムの構成

図 2 に本システムの概要を示す．システムはプロキシサーバとして実装した．以下，図 2 における各モジュール（Web ページ解析部，先読み部，要約生成部，マネージャ部）について説明する．

2.2 Web ページ解析部

このモジュールは，ユーザから要求のあった Web ページを解析し，すべてのリンク先のページの中から要約を生成するページを決定する．本研究では，リンク先ページがテキストファイルの場合のみに要約を生成する．リンク先のページがテキストファイルであるかを判定するためには，URL を手がかりとする方法もあるが，URL が.html や/などで終わるページはテキストファイルであるとみなせる．しかし，それだけでは不十分である．例えば，Yahoo ニュースでは，ニュースのリンク先が

http://headlines.yahoo.co.jp/hl?a=2
0011205-00000101-yom-soci

のようになっており，拡張子だけではテキストファイルかどうかを判定できない．したがって，実際にリンク先ページのリソースを入手し，テキストファイルかどうかを判定する．テキストファイルの時は要約を作成し，それ以外の時はリンク先ページがテキストファイルではないことを伝えるメッセージを要約の代わりに表示する．また，要約を作成するリンク先ページが決まったら，そのURL とアンカーテキストのリストをマネージャ部を通して先読み部に渡す．

また，リンクにマウスを置くとツールチップを開いて要約を提示するように参照元ページを改変する．具体的には，要約対象のアンカーに onmouseover と onmouseout を追加して，要約を表示するための JavaScript が実行されるようにする．2.1 節で述べたように，要約の生成に時間がかかる可能性があるため，参照元のページとリンク先の要約は別ページとして取り扱う．

2.3 先読み部

このモジュールでは，マネージャ部から先読み要約対象の URL とアンカーテキストのリストを受け取る．そして，これらの URL の要約を生成するように要約生成部に依頼し，生成された要約を「キャッシュ」に格納する．先読みはネットワークに対して多大なトラフィックを発生させるため，同時に先読みする数を制限する必要がある．そこで，先読みする順序を決定し，その順序に従って逐次的に先読みを行う．本研究では，先読みの順序をユーザの興味（ユーザプロファイル）とアンカーテキストから決定する．ユーザにとって興味のある単語がアンカーテキストにより多く含まれていれば，ユーザは，他のリンクと比べて，そのアンカーのリンク先ページを見たいであろうと考え，先読みの優先順位を高くする．

本研究では，ユーザプロファイルを，ユーザにとって興味があると思われる単語のリストとする．ユーザプロファイルをユーザに作成させる方法は，ユーザにかかる負担が大きい．従って，本システムでは，ユーザプロファイルは自動的に作成する．ユーザプロファイルにユーザが興味があると思われる単語を登録する方法は以下の通りである．

1. ユーザが閲覧した Web ページの HTML タグを取り除き，形態素解析にかける．
2. 形態素解析した結果の名詞，未定義語をユーザプロファイルに登録する．また，それらの出現頻度も記録する．

次に，要約対象の先読み順序を決定する方法を以下に示す．

1. アンカーテキストを形態素解析して，名詞と未定義語を取り出す．
2. 取り出した単語のユーザプロファイルにおける出現頻度の合計をそのアンカーのスコアとする．
3. すべてのアンカーについてスコアを計算し，スコアの高い順に並べて先読みの順序とする．

2.4 要約生成部

このモジュールでは要約を生成する．本研究の目的は，要約を表示することにより，リンク先がユーザにとって重要かどうかを判断する手助けをすることにあ

る．しかし，要約作成にあまりに時間がかかっては，この目的は達成されない．したがって，要約作成アルゴリズムには要約作成時間が短いことが要求される．

本研究では，要約の対象はWebページなので，それに特化した要約手法を用いることも考えられる．Webページに特化した自動要約としては，HTMLのタグや位置情報を利用した要約が考えられる．しかし，実際にインターネット上にあるWebページの多くは，HTMLのタグをレイアウトに利用しており，必ずしもドキュメントの構造をうまく記述していない．また，特にポータルサイトなどでは，広告や決まり文句（例えば，ニュースサイトでは，地方サイトへのリンクなど）が，テキスト中の様々な位置に出現し，位置情報を手がかりとした要約手法も簡単には見出せそうにない．

そこで，本研究では，Webページに特化した自動要約を考えるのではなく，テキストを対象とした既存の要約技術を利用する．具体的には，テキスト簡易要約器 Posum[2] [4] を利用し，tfに基づく重要文抽出法によって要約を作成した．

2.5 マネージャ部

このモジュールでは，Webブラウザや他のモジュールとの間の情報の受け渡しを行う．まず，Webブラウザから，Webページや要約の要求を受け付ける．通常のWebページの要求であれば，Webページ解析部に処理を依頼する．要約の要求であれば，要約を「キャッシュ」から取り出す．要約がまだ作成されていなければ，要約生成部に処理を依頼する．そして，結果のコンテンツをユーザのブラウザに渡す．また，Webページ解析部から受け取った先読み要約対象のURLとアンカーテキストのリストを先読み部に渡し，先読み処理を依頼する役割も果たす．

3 評価実験

本研究で提案するシステムの有効性を確認するために，プロトタイプシステムを用いた評価実験を行った．WWW情報探索のタスクとして，ニュースサイトを一通り閲覧し，被験者にとっての重大ニュース3つを選んでもらうタスクと，検索サイトを利用して調べ物をするタスクを実行してもらい，システムの使いやすさに関するアンケートに答えてもらった．

現在のシステムの実装は十分に効率化されていないため，先読みが完了する前にユーザがリンク先の要約を表示させることも多い．このため，先読みを行う効果に対する正当な評価が得られない可能性もある．そこで，現在のシステム（システムAとする）と，リンク先のすべての要約がキャッシュに格納されている理想的な状態を擬似的に実現したシステム（システムBとする）の2つで実験を行った．ただし，被験者には，2つのシステムの違いを説明していない．また，要約の表示方法として，別ウィンドウに表示する方法も実装して被験者に試してもらった．被験者は，本大学院の学生，卒業生の8名である．アンケートの項目と結果を以下に示す．

1. このシステムは使いやすいと思いますか

使いやすいと思う	1人
改良されれば使いやすいと思う	7人
どちらともいえない	0人
使いやすくないと思う	0人

2. 表示方法は別にして，リンク先の要約を見ることは役に立つと思いますか．

役に立つと思う	4人
要約の質が良ければ役に立つと思う	4人
あまり役に立たないと思う	0人
ない方が良い	0人

3. 要約の表示方法（ツールチップ表示）は良いと思いますか．

良いと思う	3人
表示の大きさ等が改善されれば良いと思う	4人
どちらともいえない	1人
良くないと思う	0人

4. 要約の表示方法は，ツールチップ表示と別ウィンドウ表示のどちらが良いと思いますか．

ツールチップ表示の方がいいと思う	2人
どちらともいえない	1人
別ウィンドウ表示の方がいいと思う	5人
どちらも良くないと思う	0人

5. システムAの場合，要約の表示はどう感じましたか．

遅くは感じなかった	0人
少し遅いが許容範囲だ	1人
時々遅いときがある	6人
遅いときの方が多い	1人
すべて遅い	0人

6. システムBの場合，要約の表示はどう感じまし

[2] http://nlp-www.jaist.ac.jp:8000/~motizuki/software/posumcl/

たか.

遅くは感じなかった	5人
少し遅いが許容範囲だ	2人
時々遅いときがある	1人
遅いときの方が多い	0人
すべて遅い	0人

4 考察

アンケート項目2では，リンク先の要約を見ることは,「質が良ければ」という回答も含めて，全員から役に立つという意見が得られた．このことから，要約の表示はWWW情報探索支援に有効であると言える．

本研究では，2.1項で述べたような理由から，要約の表示方法としてツールチップを用いることが有効であると考えた．しかし，アンケート項目4によれば，ツールチップ表示は「目がちらつく」といった意見があり，どちらかというと別ウィンドウ表示の方が良いという結果が得られた．しかし，アンケート項目3の回答を見ると，ツールチップ表示に対する好意的な意見もあることがわかる．したがって，要約の提示方法は，ユーザに選ばせるのが良いだろう．

アンケート項目5と6の回答を見ると，2つのシステムの表示速度には明確な差が見られる．このことは，本システムが効率化という面で十分に洗練されていないことを示唆している．先読みの効果を定量的に評価するために，システムAについて，リンク先の要約を表示する前に先読みが完了していた割合を調べた．8人の被験者の平均は27.6%であり，最低で3.6%，最高で51.0%であった．したがって，この割合を100%に近づけるようにシステムを高速化する必要がある．しかし，システムAについても,「時々遅いときがある」と答えた人が6人と最も多く，先読みの効果はある程度認められる．

5 関連研究

Webブラウジングにおいて，リンク先の情報を表示することによって情報探索支援を行う先行研究としては，以下のようなものがある．

Haraldらは，Webページ上のリンクにマウスカーソルを置くと，リンク先の様々な情報を表示するシステムを提案している[5]．表示する情報はリンク先のタイトル，言語，最終訪問日，サーバの応答速度，サイズなどである．これに対し，本研究では，このようなリンク先の情報ではなく，コンテンツの要約を表示する点が異なる．また，Haraldらは，リンク先の情報が実際にWeb探索に有用であるかを実験的に確認したわけではない．本研究では，プロトタイプシステムをを作成して評価実験を行い，要約を事前に表示させることがWeb探索の支援に有効であるとの見通しを得た．

Kopetzkyらは，Webページ上のリンクにマウスカーソルを置くと，リンク先のサムネイル画像を表示するシステムを提案している[6]．このシステムはProxyサーバで実装され，ユーザからリクエストのあったWebページを解析し，リンク先のサムネイル画像を表示できるようにWebページを改変する．しかし，サムネイル画像を作成し表示する時間的な問題に関しては全く考慮されていない．これに対し，本研究では，サムネイル画像ではなく要約を表示する，要約を先読みことにより，要約を表示する時間を短縮させている．

6 おわりに

本研究では，WWW情報探索支援のために，リンク先のページを先読みして要約を作成し，ユーザに提示するシステムを作成した．評価実験の結果，要約を事前に表示することがWWW情報探索支援に効果があることを確認した．

今後の課題として，まずシステムの効率化が挙げられる．ユーザがリンク先のページを閲覧しようとする前に常に要約の作成を終えるようにしなければならない．また，現時点では，要約の作成に要する時間も無視できない．要約の質を向上させることも重要だが，高速に要約を生成する手法を開発することも重要な課題である．

参考文献

[1] Linda Tauscher and Saul Greenberg. How People Revisit Web Pages: Empirical Findings and Implications for the Design of History Systems. International Journal of Human-Computer Studies, 47(1):97 - 138, 1997.

[2] 奥村学, 望月源,「テキストを自動的に要約する技術 -第1回- テキスト中の重要な文を抜き出す」, コンピュータサイエンス誌 bit 2月号, 共立出版, pp.37-42,2000.2.

[3] 奥村 学, 難波英嗣 ,「テキスト自動要約に関する研究動向」, 自然言語処理, Vol.6, No.6, pp.1-26, 1999.

[4] 望月源,「テキスト簡易要約器 Posum version1.50.2 マニュアル」, JAIST Technical Memorandum, IS-TM-2002-002.

[5] Harald W. R. Weinreich and Winfried Lamersdorf. Concepts for Improved Visualization of Web Link Attributes. In Proceedings of the WWW9, 2000.

[6] Kopetzky, T. and Muhlhauser, M. Visual Preview for Link Traversal on the WWW. In Proc. 8th Intl. WWW Conf.,May 1999, 447 - 454.