マルチメディア通信と分散処理　92－5
（１９９９．２．２５）

ネット情報自動ガイド技術を用いた情報検索支援システムの開発

林　憲亨　　新井　克也
NTT 情報流通プラットフォーム研究所

インターネット上から必要な情報を含む Web ページを探し出す手段として、検索サービスが多く利用されている。その中でロボット型検索サービスは検索する Web ページのテキストを膨大なデータベースに保存しており、ディレクトリ型検索サービスでは見つけることができなかった情報も、ロボット型検索サービスを利用することにより見つけることが可能となる。しかし、ロボット型検索サービスは検索結果が膨大な数になることが多く、その中から必要な情報を探し出すまでに多くの時間を費やしてしまう。

今回、ネット情報自動ガイド技術を用いた情報検索支援システムを開発した。このシステムは既存のロボット型検索サービスをそのまま利用し、単位時間当たりに閲覧する Web ページ数を増加させることにより必要な情報を含む Web ページを探し出すまでの時間を短縮させることが可能である。本稿では、この情報検索支援システムについて報告する。

Development of Information Retrieval Support System using Net Information Auto Guide Technology

Noriyuki Hayashi and Katsuya Arai
NTT Information Sharing Platform Laboratories

When we retrieve Web including necessary information for us on Internet, we use Search Engines. Robot-driven Search Engine have a lot of information data, so it is possible that the data include necessary information of user. However, Robot-driven Search Engine give us a huge number of retrieval results. It takes a lot of time to retrieve necessary information in the retrieval results.

We developed information retrieval support system using net information auto guide technology. This system improve the problem of information retrieval work from a different point of view. In this paper, we report this system.

1．まえがき

現在、インターネット上には 3 億以上の Web ページが存在すると推定されており[1]、その中から必要な情報を含む Web ページを探し出す手段として、検索サービスが多く利用されている。その中でロボット型検索サービスは Web ページのテキストを膨大なデータベースに保存しており、ディレクトリ型検索サービスでは見つけることができない情報も、ロボット型検索サービスを利用することにより見つけることが可能となる。しかし、ロボット型検索サービスは検索結果が膨大な数になることが多く、その中から必要な情報を探し出すまでに多くの時間を費やしてしまう。

今回、ネット情報自動ガイド技術を適用した情報検索支援システムを開発した。本システムは既存のロボット型検索サービスをそのまま利用し、単位時間当たりに閲覧できる Web ページ数を増加させることにより、必要な情報を含む Web ページに辿り着くまでの時間を短縮することが可能とする。本稿では、この情報検索支援システムの構成を述べ、評価モデルを用いた検索結果閲覧速度の検証結果について報告する。

2．情報検索の効率化

効率の良い情報検索とは利用者が必要な情報をできるだけ短い時間で探し出せることである。従って、ロボット型検索サービスを用いて検索を行なう場合、利用者が必要な情報を含む Web ページが一番初めに提示されるのが最も効率の良い情報検索となる。現在運営されているロボット型検索サービスは各社独自の方式で順位付けを行ない、利用者の必要な情報を上位に提示する試みが行なわれている[2][3][4][5]。しかし、利用者から入力されたキーワードだけで利用者の必要な情報を含む Web ページの URL を上位に位置付けるのは難しい。従って、検索結果のリスト内のどこに必要な情報を含む Web ページ URL が記述されているかわからないため、提示された Web ページの URL リストを多く確認する必要が生じる。

ロボット型検索サービスは検索結果 Web ページの URL リストと一緒にそのリスト先ページの概略を併記している。しかし、その概略はリスト先ページの先頭の文章を抜き出しているだけであるため、概略だけでそのページに必要な情報が含まれているかどうか判断することは難しい。従って、リストと概略だけではなく、リストの Web ページの内容も確認する必要が生じる。

従って、従来のロボット型検索サービスの利用において、必要な情報を含むページを探し出すまでに時間がかかった要素は以下の 2 つである。

①提示された Web ページ URL のリストを多く確認する必要がある

②提示された Web ページ URL の内容を確認する必要がある。

この２つの要素をできるだけ短い時間で行なうことにより、効率の良い検索が可能となる。

3．システムの提案

（1）提案する方式

従来のロボット型検索サービスを用いて、効率の良い検索を行なうためには、

①提示された Web ページ URL のリストを多く確認する必要がある。

②提示された Web ページ URL の内容を確認する必要がある

の 2 つをできるだけ短い時間で行なうことが重要である。そこで本システムでは従来のロ

ボット型検索サービスを利用し、単位時間当たりに多くの検索結果そのものの Web ページを閲覧可能にする方式を検討する。すなわち「単位時間当たりに多くの Web ページを表示する」ことを目的とする。多くの Web ページを表示するために、空間軸と時間軸を有効利用する。空間軸とは端末画面、時間軸とは表示時間、表示待ち時間のことである。本方式はロボット型検索サービスから出力される検索結果のリストを複数項目ずつ、Web ページそのものを同時に表示する。同時表示数、表示時間はネットワークの速度、端末画面サイズから決定する。本システムはロボット型検索サービスの検索結果として利用者の必要な情報がリストの 1 番初めに提示されていた場合、検索にかかる時間は従来と同じになるが、2 番目以降に提示されていた場合は検索にかかる時間を短縮することが可能となる。従って、本システムを利用することにより検索の効率が悪くなることはない。また、Web ページを閲覧し、その内容を確認するためには Web ページの同時表示数の最大値、表示時間の最小値が存在し、この値を超えると内容の確認が不可能となる。Web ページの同時表示数については 8 ページ程度までは同時確認可能であり、それ以上を超えると確認能力が低下するとの実験結果が得られている[6]。また端末画面サイズから SVGA（800×600）サイズで同時表示数 4、XGA（1024×768）サイズで同時表示数 9 が限界であり、これ以上は 1 つの Web ページの表示領域が小さくなり、内容の確認が困難となる。また図 1 に示すように同時表示数 9 ページの確認には経験的に最低 10 秒が必要である。すなわち表示時間と表示待ち時間を合わせた表示間隔を 10 秒にすることがシステムとして理想である（図 1）。

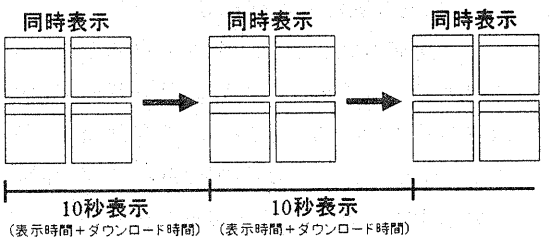

図１：理想的な表示の流れ

（2）システム実現における問題点

本システムを実現するための解決すべき問題点として 3 つ考えられる。

① 操作上の問題

検索結果として表示されるのは URL のリストである。従って、情報を取得するためにはそのリストをマウスクリックする必要があり、リンク先の Web ページを閲覧し、その内容を確認し、必要とする情報を含まないときはまた URL のリストのページに戻って、同じ作業を繰り返す必要があった（図 2）。

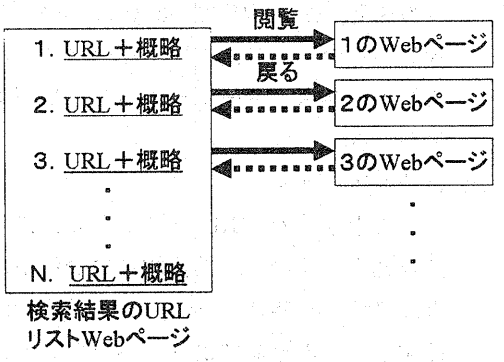

図 2：従来の閲覧

② 表示待ち時間の問題

ネットワーク環境、サーバの反応速度により、検索結果の URL をマウスクリックしてから Web ページが表示されるまでに Web ページデータをダウンロードするまでの表示待ち時間が生じる。

③ 逐次的な閲覧による問題

従来のブラウジングはマウスクリックで進めていくため、１つのブラウザ単位での確認が基本である。しかし、端末画面サイズの高精細化により、複数の Web ページを同時に表示することが可能となってきている。このような状況にもかかわらず、従来の１ブラウザ単位での逐次的な閲覧では端末画面サイズを有効に利用するのが難しい。

（３）解決に必要な機能

本システムを実現する上で実装すべき機能は以下の３つである。

①Web ページ自動表示機能

これまではマウスクリックによって Web ページの表示を行なっていたが、検索結果として URL リストを利用し、結果の Web ページそのものの提示を自動化する機能が必要である。これによりマウスクリックによる検索結果の Web ページへの移動、および検索結果 Web ページ閲覧後、そのページに必要がない場合に検索結果のリストページへの戻る作業が不要となる。

②先行読み込み機能

検索結果の URL リストを元に先行読み込みを行なう。具体的には表示中に次に表示する Web ページの読み込みを開始する。この機能により、ネットワークを有効に利用し、表示待ち時間を最小限にすることが可能となる。

③ブラウザ表示制御機能

ブラウザの表示サイズ、表示位置、表示時間の制御により、端末画面中に同時に複数の Web ページの表示を行なう。この機能により、人間の識別能力を活かすことができ、短時間に多くの Web ページの確認が可能となる。

４．情報検索支援システム

（１）ネット情報自動ガイド技術

今回開発したシステムはネット情報自動ガイド技術（アプリケーション名：Web 紙芝居[7]）を利用した。Web 紙芝居はＮＴＴで開発した技術であり、前章で述べた機能を全て網羅している[8]。

Web 紙芝居はマルチメディアリソースの同期を制御した表示が可能なシステムである。マルチメディアリソースとして文字、音声、画像、Web ページの利用が可能である。次に Web 紙芝居の実行時の流れについて説明する。クライアント側端末に Web 紙芝居実行エンジンが設定されてない場合、サーバから自動的にダウンロードされる。Web 紙芝居は実行開始時にサーバ上の表示制御コマンドファイルを読み込み、ファイルに記述されたコマンドに従って実行される。Web 紙芝居の実行の流れを図３に示す。

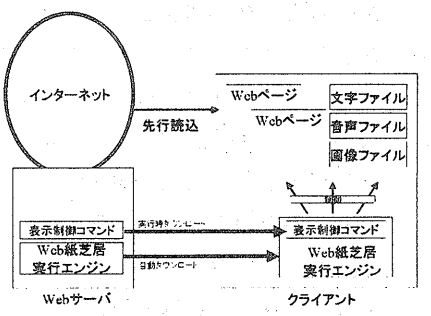

図３：Web 紙芝居実行の流れ

前章で述べた機能と関連する Web 紙芝居の技術的特徴を以下に述べる。

①キャッシュフィードバック技術

ネットワーク性能を考慮した表示の高速化を行なう。これは先行読み込み機能により次の表示に必要なコンテン

ツ（Web ページ、音声ファイル、画像ファイル）を読み込み、ネットワークを有効利用することにより表示待ち時間の最小化が可能となる。また、ネットワーク性能に応じたマルチメディアリソースの自動選択を行なう。

②ブラウザ表示制御技術

クライアント側の端末画面性能に応じたブラウザの表示順序・表示位置・表示サイズ・表示時間の制御を行なう。

（2）検索出力変換サーバ

Web 紙芝居をユーザインターフェースとして利用するためには、利用者の希望に応じた Web ページの表示制御コマンドを自動的に生成するサーバが必要となる。今回開発した検索出力変換サーバは利用者からの要求として、ブラウザ数、ブラウザの大きさ（画面の大きさ）、表示時間などの表示系パラメータ、および、通常の検索サービスの利用時に入力するキーワード、検索結果数などの検索系パラメータを受け付ける。検索出力変換サーバは通常の検索系パラメータを検索サービスに入力し、検索結果の URL リストのページをもらう。検索結果の URL リストのページから検索結果の URL を抜き出し、表示系パラメータをコマンドとし組み込んだ表示制御コマンドを生成する。Web 紙芝居と検索出力変換サーバを組み合わせたシステムを図 4 に示す。

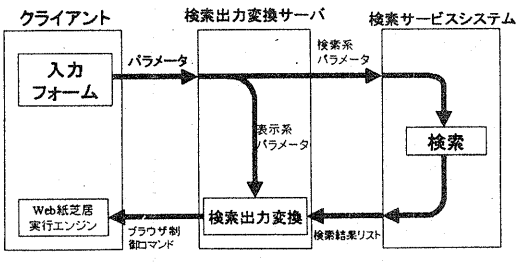

図４：本システムの構成

５．本システムによる検索結果閲覧速度の高速化

従来の検索と開発したシステムの評価モデルを用いて検証を行なった。評価モデルとして用いた Web ページ閲覧モデルを図 5 に示す。

図中の記号 T_{dl}、T_{br} はそれぞれ次の Web ページ表示に必要なダウンロード時間、端末画面中に表示された Web ページ全部の確認にかかる時間である。また N は端末画面中の同時表示 Web ページ数、M×N は総閲覧ページ数である。

この図より、従来のブラウジングで M×N ページを見る場合に要する時間は

$M \times N \times (T_{dl} + T_{br})$

また、本システムを利用した場合のブラウジングで M×N ページの閲覧に要する時間は

$T_{dl} \leqq T_{br}$ の場合　　$T_{dl} + M \times T_{br}$

$T_{dl} \geqq T_{br}$ の場合　　$T_{dl} + M \times T_{dl}$

となる。

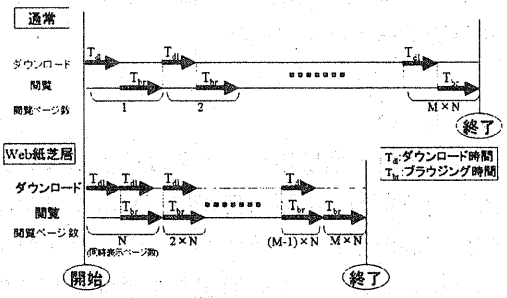

図５：閲覧モデル

図 6 は T_{br} を 10 秒（理想値）、Web ページの平均データ量を 50KB、理想コンピュータ、理想ネットワークとした場合の検索結果閲覧速度のグラフである。同時表示ブラウザ数は通常のインターネット閲覧時に使用

するブラウザサイズを考慮して SVGA（800×600）のとき 4 ブラウザ、XGA（1024×768）のとき 9 ブラウザとした。検索結果閲覧速度とは通常のブラウジング（逐次手動型閲覧）と比較した単位時間内に閲覧できるページ数比のことである。また図中の○は実験値である。

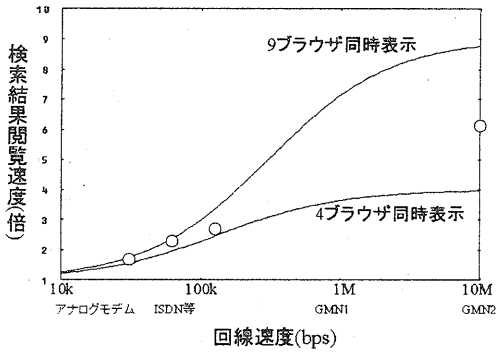

図６：検索結果閲覧速度

このグラフより、従来に比べ短時間で多くの Web ページの確認ができることがわかる。特にイントラネットで利用されている LAN 回線（10Mbps）においては、端末画面サイズが XGA（1024×768）のとき、実験値で従来の 6 倍、理論値では従来の 9 倍の検索結果閲覧速度が得られた。

６．まとめ

ロボット型検索サービスは検索結果が莫大となるため、必要な情報に辿り着くまでに時間がかかるという問題があった。今回開発したシステムは、この問題点を単位時間当たりに閲覧する Web ページ数の増加による解決を目的としている。評価モデルを用いて本システムの検証を行なった結果、実験値で従来の 6 倍、理論値で従来の 9 倍の検索結果閲覧速度が得られ、本システムは検索効率の改善に効果があることを示した。

参考文献

[1] Lawrence and Giles : "Searching the World Wide Web", Science, vol.280. pp.98-100(1998)

[2] http://www.goo.ne.jp/help/faq2.html

[3] http://www.lycos.co.jp/help/result_help.html

[4] http://www.excite.co.jp/more_excite/search_1_2.dcg

[5] http://navi.ntt.co.jp/faq/

[6] 溝渕、古山："WWW の情報取捨選択行動におけるインデックスの影響分析"、情報処理学会研究会 HI81-6、1998 年 12 月

[7] Web 紙芝居、商願平 9-109133、9-109134

[8] 林、新井、西田："自動ブラウジングシステムの検索サービスへの適用"、情報処理学会第 55 回全国大会、1997 年 9 月