

## リンク情報を利用したインターネット情報検索方式の検討

5 Z - 7

藤井洋一、鈴木克志

三菱電機株式会社 情報技術総合研究所 音声・言語インタフェース技術部

### 1.はじめに

近年大量のテキスト情報がインターネットを通じてアクセス可能となるにつれて、大量の情報の中から利用者の必要とする情報を的確に見つけ出す必要性が益々高くなってきている。

現在、インターネット利用者が必要とする情報を見つけ出すための方法として、情報を収集して検索する機能を提供している検索サービスを利用する方法がある。検索サービスには、AltaVista<sup>[1]</sup>に代表されるロボット(WWW ロボット)と呼ばれるソフトでページを収集し、収集ページに対して全文検索を行うロボット収集型と、Yahoo! Japan<sup>[2]</sup>に代表される利用者から登録要求があったものに対して人手によって分類・登録し、分類検索(分類をたどりながら情報にたどり着く方法)と、登録した情報の全文検索を行うディレクトリ登録型がある。これらの検索サービスには表1の問題点がある。

表1 検索サービスの問題点

ロボット収集型	検索結果が非常に多くなり、絞り込みが困難である。 ロボットでの収集に時間がかかり、最新の情報である保証がない。
ディレクトリ登録型	希望する情報が登録されていない場合がある。

一方、情報のリアルタイム性を向上させるため、特定の情報の更新を定期的にチェックに行き、更新があったことを利用者に提示するオートパイロットソフトウェアがあるが、対象とする情報(URL)を利用者が前もって知っていなくてはならない。

今回、インターネット上の情報の更新チェックを定期的に行い、さらにリンク情報を元に利用者にとって未知の情報も提示する情報検索方式を検討した。

### 2.従来の方式について

インターネット上の情報を利用する目的は、(1)新しい情報を入手するためと、(2)既にある情報が更新された時に更新情報を入手するための2つに大きく分けられる。(1)は、一度検索をして結果が分かれば良く、(2)は、特定の情報を監視し更新をチェックする必要がある。

従来のソフトウェアを利用して、利用者が情報を検索・監視するための方法としては、(1)では、検索サービスで検索した結果を眺め必要な情報を見つけ出す方法が考えられるが、実際には絞り込むことが非常に困難である。

一方、(2)では、図1のように検索サービスで検索した結果の中から関係する情報(URL)を見つけ出し、その情報をオートパイロットソフトで監視するという方法が考えられるが、利用者にとって未知である情報を監視することはできない。

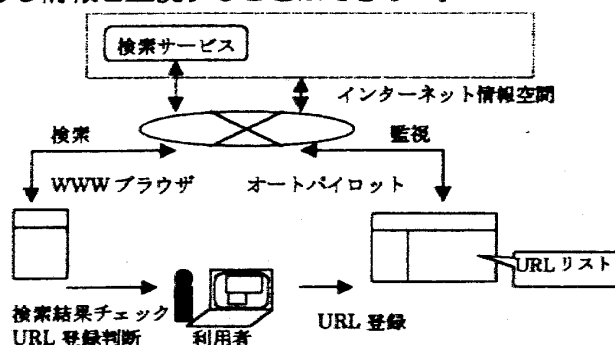


図1 従来の方式

### 3.本検討の方式について

本検討の方式は利用者の検索意図にあった検索・監視を実行するために、図2のようにあらかじめ利用者に既知である情報で検索して検索条件を詳細化しておき、その検索条件に従ってリンクを辿りながら利用者の検索条件に一致するWWWページを提示するとともに、収集対象に追加し監視する。この時、リンクの辿り方を工夫することでより速く

A study of information retrieval method on the internet using link information  
 Youichi Fujii, Katsushi Suzuki  
 Human Media Technology Dept. Information Technology  
 R&D Center, Mitsubishi Electric Corp.  
 5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, Japan

利用者が希望する情報を収集対象とする。

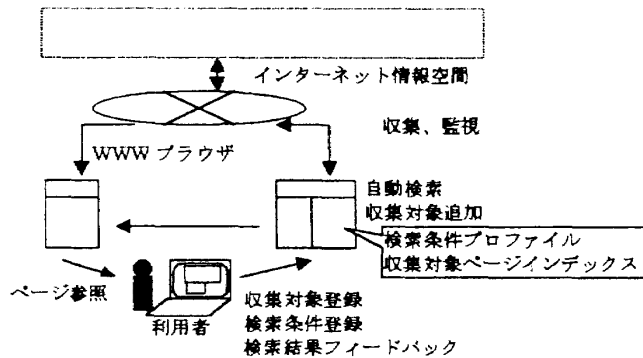


図2 本検討の方式

具体的な処理の流れを示す。

1. 利用者はあらかじめ監視したい収集対象を登録。
2. 利用者の指定にしたがって本方式ではページを収集し、ページ中の単語や、リンク間の情報等を保存。収集・更新作業は定期的に行う。
3. 利用者は検索したい内容の検索条件を登録。
4. 検索条件に従って検索を実行し、利用者の関連するしないの検索結果フィードバックを元にベクトル表現した検索条件プロファイルを作成。
5. 定期的に検索結果のリンク間の情報を元に収集対象候補を抽出し、各収集対象候補に対して優先的に仮収集したページと検索条件プロファイルとの類似度計算で、追加収集対象とするかどうかを判定、1.の収集対象に自動的に追加する。

3.1. 収集対象候補抽出方式

収集対象候補抽出方式は、関連情報収集型と詳細情報収集型の2つを考える。関連情報収集型はリンク集的なページに関連する情報が記載されていることを利用する。また、詳細情報収集型は、検索結果の特定内容の詳細な内容説明のリンクを抽出する。

(1) 関連情報収集型

関連情報収集型は、図3のように他のサイトのページを集めたリンクページが検索結果であり、その幾つかのリンク先も検索結果である場合、リストで併記された未収集のページを優先的な収集対象候補とする。

関連情報収集型は、利用者が知らないうちに新たに追加されたような情報や、利用者にとって未知の関連度が高い情報が見つかる効果がある。

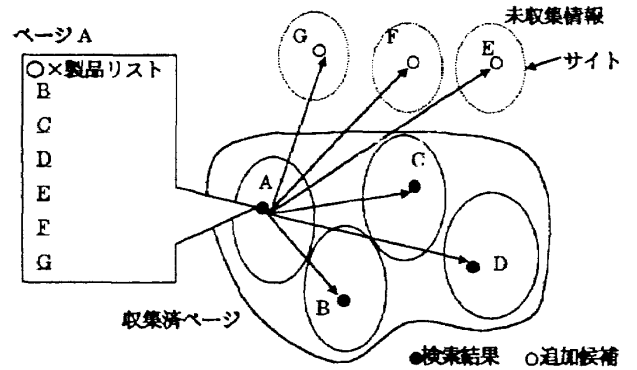


図3 関連情報収集型

(2) 詳細情報収集型

詳細情報収集型は、図4のように検索結果中の複数のページから未収集の同じページを提示しているものを優先的な収集対象候補とする。

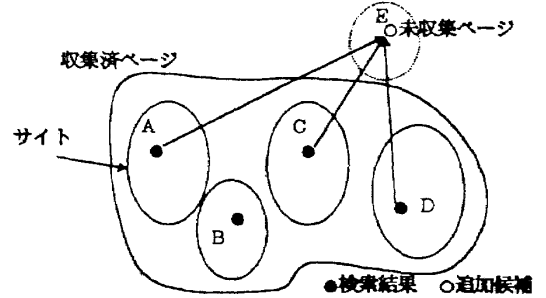


図4 詳細情報収集型

詳細情報収集型は、昔から存在する情報ではあるがこれまで利用者にとっては未知であった情報が見つかる効果がある。

4. まとめ

今回、リンク情報を利用したインターネット情報検索方式を検討した。検討した方式は、従来の検索サービスのように利用者が大量の検索結果から情報を探し出す必要がない。

インターネット上の情報を利用者の要求にしたがって収集し、かつ検索条件にあった検索結果を提示する。さらに、検索条件との類似度を判定しながら収集・監視する対象を追加することで、利用者にとって未知の情報を効果的に提示できることが可能となると考えられる。

今後は、本方式の有効性を試作し、評価する。

[参考文献]

- [1] AltaVista <http://www.altavista.com>
- [2] Yahoo! Japan <http://www.yahoo.co.jp>