マルチメディア通信と分散処理　82－31
（１９９７．４．２５）

WWW情報検索を支援するバーチャル図書館システム

－ システムコンセプトと用途別検索方式 －

石川 浩通　谷田 望　徳永 寿郎　田中 聡
三菱電機株式会社　情報技術総合研究所

インターネット等の広域ネットワーク上に分散し、種々のユーザが構築した、多種多様でかつ常に更新が行われるような既存のデータベースの中から、目的の情報を探し出すのは非常に大変である。これに対して、情報検索の専門家でない利用者でも、目的とする情報の所在を容易に探し出せるように、検索を支援するバーチャル図書館システムの開発を行っている。

本稿では、現状のWWW情報検索における問題点を整理し、これらの課題を解決するためのバーチャル図書館システムの開発コンセプトと、その技術的な特長について論じた後、医療を具体対象として試作を行ったバーチャル医療図書館システムの概要と、その評価結果について述べる。

Development of Virtual Library System

－ Its concept and retrieval method based on categories －

Hiroyuki Ishikawa　Nozomu Tanida　Toshiro Tokunaga　Satoshi Tanaka
Information Tecnology R&D Center, Mitsubishi Electric Corp.

It is very difficult for users to acquire the information which they really want among various, enormous, and constantly changing data from broad network like WWW of the Internet. We are developing a Virtual Library System in order that even a user, who is not accustomed to the computer system, can find out the location of the targeted information with ease.

In this paper we discuss the problems of WWW information retrieval ,and explain the concept and technical points of a Virtual Library System in order to solve these problems.

1 はじめに

近年、インターネットでのWWW(World Wide Web)の発展はめざましく、既にネットワーク上には、広範囲でかつ膨大な量の情報が存在し、今後さらに、情報内容の多様化、情報量の増大が進むと考えられる。このような状況の中で、各種の情報検索サービスが、商用やボランティアで複数立ち上がっており、WWW情報に対する検索が可能となってきた。しかしながら、これらのサービスは不特定多数の利用者を想定した汎用的なものである。このため特定分野の詳細な情報を検索したいというような場合に、一般的なキーワードで検索すると大量の検索結果が提示され、その中から目的の情報を絞り込まなければならず、また、詳細な検索条件を指定すると、検索結果が得られないなど、目的の情報になかなか到達できないのが現状である。

これに対して、我々は、様々な検索者の様々な検索目的に柔軟に対応でき、情報検索の専門家でない利用者でも目的の情報を容易に検索できることを支援するバーチャル図書館システムの開発を行っている。

本稿では、2章において、現状のWWW情報検索における問題点を整理し、3章、4章において、これらの課題を解決するためのバーチャル図書館システムの開発コンセプトと、その技術的な特長を述べる。また、5章、6章において、医療を具体対象として試作を行ったシステムの概要と、その評価結果について述べる。

2 WWW情報検索における問題点

WWW情報検索における問題点を述べるにあたり、まずWWW情報空間における情報資源の持つ特徴について整理する。

・大規模・多様性：

インターネットの普及に伴って、誰でもが世界中に向け自由に情報を発信することが可能となった。このため、現在、インターネット上には非常に膨大で、かつ、個人情報から、専門的な情報まで多種多様な情報が蓄積されている。

・未整理：

従来のデータベースに格納される情報は、データベース設計者の設計したデータモデル（スキーマ）により整理されたものである。これに対して、WWW情報は、このような管理者によって統一されたモデルを持っておらず、ほとんど未整理のままの状態で蓄積されている。

・ダイナミック性：

WWW上の情報は、常に追加・更新・削除が発生する。しかも、これらの変更は、情報発信者の判断により独自に行われもので、利用者に対して変更を通知するような管理の仕組みは存在しない。

これに対して、WWW情報検索サービスは、利用者が目的の情報にたどり着くための唯一の支援となっている。しかしながら、現状のWWW情報検索サービスには、以下のような問題点があり、一般の利用者では目的の情報を容易に見つけ出すことは困難である。

(a)検索ノウハウが必要：

あるキーワードで検索を実行すると、膨大な量の検索結果が提示され、その中から目的の情報を絞り込まなければならないなど、検索条件の指定や絞り込みに高いノウハウが要求される。

(b)特定分野の詳細な情報を検索できない：

従来からカテゴリ検索と呼ばれる分類情報を用いた検索方式が多く採用されているが、これは一通りのカテゴリで一般的な分類分けをされており、特定分野の情報を、様々な観点から、詳細に検索したいというような検索要求には応えられない。

(c)データ更新を補足できない：

検索した結果みつかったＵＲＬが既に移動・消滅しており、情報がなくなっている場合が多々ある。

なお、上記(a)，(b)の問題点は、WWW情報の大規模・多様性，未整理性に、(c)の問題点はダイナミック性によるところが大きいと考えられる。

3 開発コンセプト

2章で述べるような課題を解決し、様々な検索者の様々な検索要求に柔軟に対応し、情報検索の専門家でない一般の利用者でも、目的の情報を容易に検索できるようにすることが我々の開発目的であり、その解決策として、バーチャル図書館システムの構築を目指している（図1）。

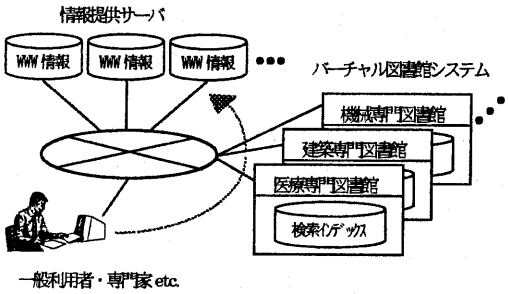

図1：バーチャル図書館システムの概念

バーチャル図書館システムとは、専門分野・メディア等に対応する用途別のWWW情報検索システムを構築し、個々の分野に特化した高度な検索支援サービスを提供するとともに、複数のシステムが協調することにより、検索者の様々な検索要求に対応できる検索システムである。利用者は、検索目的に適した用途別のWWW情報検索システムを選択することにより、ネットワーク上に分散する多種多様で、かつ大量の更新が常に行われるWWW情報の中から、目的とする最新の情報を容易に検索することができるようになる。また、インデックス更新機能を備えることにより、過去に収集した情報が変更されていないか定期的に監視し、インデックスを常に最新の状態に保つことにより、利用者は目的とする最新の情報を得ることができる。

このようなバーチャル図書館システムでは、検索ガイダンス機能や、3次元CGを用いた高度検索インタフェース等の検索支援方式を開発することを検討しているが、今回は、その検索支援の一つとして、用途別検索方式について開発を行ったので報告する。

用途別検索方式とは、収集したWWW情報に対して、複数の分類観点より検索を行える方式である（図2）。例えば、医療に関する情報を例とした場合でも、「医学分類」や「人体構造」または「病名」から検索したいなど、その検索要求は、検索者や検索目的によって異なる。また、利用者によっては、普段使い慣れている本の目次や、よく知っている分類体系より検索したい場合などがある。このように、様々な検索者の様々な検索要求に対応できるように、複数の分類観点からWWW情報を検索することを可能にしている。

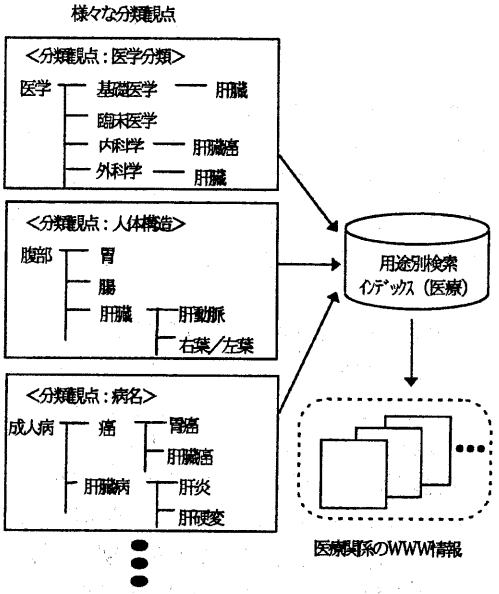

図2：用途別検索方式の検索例（医療）

このようなバーチャル図書館システムにおける用途別検索方式を実現するためには、以下のような課題を解決する必要がある。

①用途別検索方式における情報収集の課題

専門分野に対応するWWW情報検索を実現するためには、ネットワーク上に分散するWWW情報の中から、特定分野の情報だけを絞り込んで収集する必要がある。これに対して、現状、インターネットでは、一般にロボットと呼ばれるソフトウエアが存在する[1]。これらは、WWW上の情報を収集した際

に、その情報からのハイパーリンクを解析し、収集情報にリンクされた他の情報を再帰的に収集するものである。現在この種のロボットの基本的な技術は既に確立されているといってよい。しかしながら、バーチャル図書館システムを構築するためには、単に情報を再帰的に収集するだけでなく、特定分野の情報だけを絞り込んで収集する必要がある。現在、このように、特定分野の情報だけを絞り込んで収集するには人間が情報を選別する以外に良策はなく、これを自動的に行うことは難しい。

②用途別検索インデックス作成における課題

用途別検索を実現するためには、収集した情報と、電子化した複数の分類体系との対応付けを行う必要がある。しかしながら、現状のカテゴリ検索における対応付けは、ほとんど人間が行っているのが現状であり、非常に手間と労力を要している。このような対応付けを支援することはバーチャル図書館システムを構築する上で重要な課題である。

③検索インデックス更新における課題

WWWでは常に情報の変更が発生するため、一度情報を収集した後に、もとの情報に変更が発生していないかチェックし、変更があった場合には、再度情報を収集し、検索インデックスの内容を更新していく必要がある。しかしながら、WWW情報の変更は、情報発信者によって任意の時点で独自に行われ、かつ通信異常などネットワーク上の問題もあるため、このような更新を効率的に行うことは難しい。

4 開発技術

4．1 情報収集支援技術

特定分野の情報だけを絞り込んで収集するためには、情報の収集範囲を何らかの方法で制限することが必要である。このような情報の絞り込みを行う方式としては、対象となる情報の内容を表すキーワード等の内容情報をもとに、目的とする情報のみを自動で絞り込んで収集する方式[2]がある。しかしながら、この方式では、オントロジーなど対象とする分野に関連する知識を、各分野毎に構築しておかなければならないという問題がある。これに対して、我々は、WWW情報がお互いに関連を持ったハイパーメディアになっている点に着目し、URLアドレスによって、収集範囲を限定するという方式を採用した[3]。図3は本方式による情報収集の例を示す図である。具体的には、以下の2つの方法で収集範囲を限定する。

・ドメイン・パス制限：

収集開始URL(Root URL)を幾つか指定すると、そこからリンクされたURLを収集するが、収集範囲をRoot URLと同じドメイン名・パス名をもつものに制限する。ここでいうドメインとは、あるマシンのIPアドレス又はそれにつけられた名称と、通信の際のプロトコル名を意味する。例えば、http://A/B/C/x.html なるアドレスがあったとすると、http://A/までがドメイン名である。またパスとは、そのマシンでのWWWサービスのルートディレクトリからのパスのことであり、この例ではhttp://A/以下の/B/C/がパス名となる。収集する対象をRoot URLと同じドメイン・パスを持つURLのみに制限するのは強力な収集範囲指定方法であり、これによってユーザが意図しないURLを収集することを防ぐことが容易に行える。

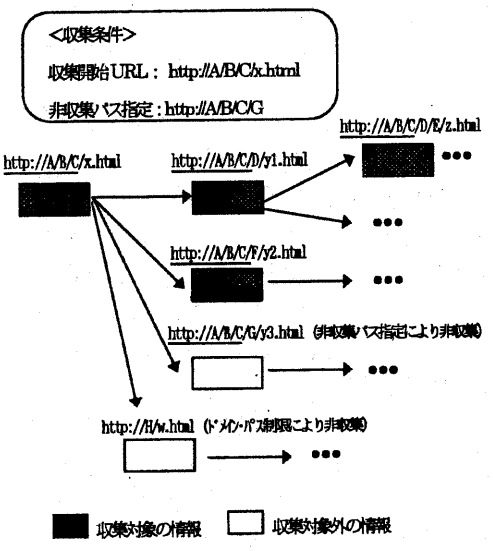

図3：情報収集の例

・非収集パスの設定：

ユーザが収集した範囲の情報を見て、不要と思えば収集範囲から外すことができる。これは収集開始ＵＲＬごとに、非収集パスというものを設け、そのパスに指定された文字列をＵＲＬ中に持つものは収集対象外とすることで、任意のＵＲＬを収集範囲から外せることにより実現した。

上記の方式は、自動で絞り込みを行うことを目指したものではなく、人間とシステムが協調することにより半自動で絞り込みを行えることを目指したものである。このような半自動的なアプローチは、人間による支援を必要とするものの、予め対象とする分野に関連する知識を整理する必要はなく、多種多様な情報を蓄積するWWW情報の収集を目的とした場合、非常に有効な方法であると考えられる。

４．２　用途別インデックス作成支援技術

用途別検索方式を実現するためには、収集した情報と、電子化された分類情報との対応付けを行う必要がある。これに関し、テキスト情報の自動分類に関する研究[2][4]が行われているが、これらの方式では、分類辞書など対象とする分野に関連する知識を、各分野毎に構築しておかなければならないという問題がある。これに対して、我々は、情報収集の場合と同様の理由で、人間とシステムが協調することにより半自動で対応付けを行う方式を採用した[5]。図４は本方式における対応付けの流れを示す図である。

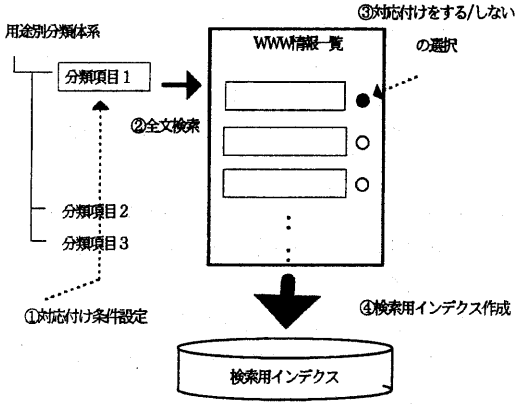

図４：分類体系とWWW情報との対応付け

図において、まず、分類を行う分類項目を選択し、その分類項目に関連のあるキーワードを設定する（対応付け条件設定）と、設定されたキーワードによってWWW情報の全文検索が実行され、検索結果として対応付け可能なWWW情報が一覧表示される。この結果を参照し、WWW情報毎に本分類項目への対応付けをする／しないを選択することにより、分類作業を行うことができる。

４．３　検索インデックス更新技術

更新収集を効率的に行うには過去の収集結果をできる限り利用することが有効であると考える。このため、今回、以下のような更新方式をとった。

・前回収集時刻から設定した収集更新間隔を過ぎていないURLは更新有無等を調べず、収集もしない。
・リクエストメソッドとしてHEADを用いて情報の更新時刻取得を行い、サーバ更新時刻の変更されていないものは収集しない。
・サーバ更新時刻が比較できない場合にはGETメソッドで収集を行うが、前回収集時に対してファイルサイズのかわっていないものは更新されていないと判断し、検索インデックスを作り直さない。

更に、更新収集間隔はユーザが任意に決めることができ、サーバによるデータの更新率を考慮しながら決定することにより、効率的な収集が可能となる。

５．バーチャル医療図書館システムの概要

上記で開発した技術を基に、医療を具体対象としたバーチャル医療図書館を開発した。図５にシステムの構成を示す。本システムは、以下の機能から構成されている。

①WWW情報収集機能

本機能は、インターネット上を定期的に巡回し、検索対象となるWWW情報の収集を行う機能である。収集したWWW情報はシステム内に蓄積し、検索インデックスの生成に利用される。また、本機能は収集したWWW情報を定期的に監視し、常に最新の情報を収集する。

②用途検索インデックス作成支援機能

本機能は、WWW情報収集機能にて収集したWWW情報から検索用インデックスを生成するシステム管理者の作業を支援する機能である。

③用途別検索機能

本機能は、WWW情報検索のためのガイダンス機能である。 検索者は複数の分類体系の中から、自分の検索目的に応じた分類体系を選択することにより、効率的にWWW情報検索を行える。

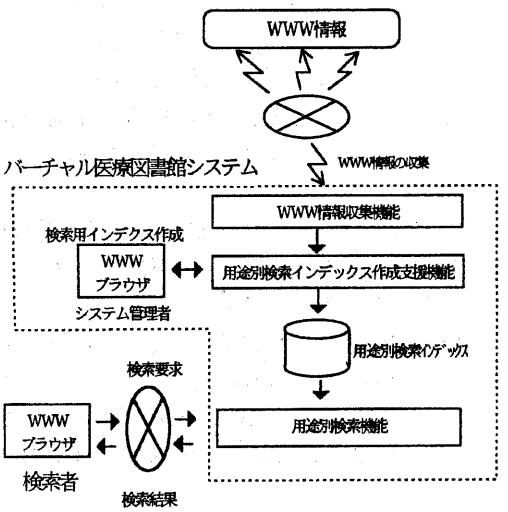

図5：バーチャル医療図書館システムの構成

6．評価

（1）情報収集

医療関係の6つのサーバのトップアドレスを収集開始URLとし、ドメイン・パス等の収集範囲制限機能により、医療に関係する情報のみを収集することができた。初期収集量は30,000URL程度で容量総計は約80Mbyteである。初期収集時のダウンロード速度は平均1.27URL/sec (3.00kbyte/sec)で、総収集時間は6.6時間という結果であった。なお、今後の課題としては、収集範囲の設定をよりきめ細かく行う必要がある場合に、非収集パスの設定を支援する機能の開発があげられる。

（2）用途別検索インデックス作成

上記で収集したWWW情報のうち、8000URLに対し、約532項目から構成される分類情報との対応付けを行った結果、本機能を用いて検索インデックスの構築を行う場合、自動対応付け、及び対応付け結果の確認／修正作業に要する時間を合わせ、約45時間を要した。これに対し、人間がすべて手作業で対応付けを行うのに要する時間は、机上見積りで約130時間かかると予測される。この結果、本方式により約1/3の効率化が図られたと考えられる。

7．まとめ

インターネット等の広域ネットワーク上に分散し、種々のユーザが構築した、多種多様でかつ常に更新が行われるような既存のデータベースの中から、情報検索の専門家でない利用者でも、目的とする情報の所在を容易に探し出せるように、検索を支援するバーチャル図書館システムのコンセプトと、バーチャル図書館システムを構築するための用途別検索方式についての開発内容を報告した。

今後は、検索インタフェースの高度化や、検索ガイダンス機能等の開発を行っていく予定である。

参考文献

[1]M.Koster:"Guidelines for Robot Writers," http://info.webcrawler.com/mak/projects/robots/guidelines.html.

[2]岩爪,武田,西田：弱構造化オントロジーを用いたインターネットからの情報獲得,電子情報通信学会技術研究報告,No. AI95-32, pp.63-70,1995.

[3]谷田,石川,須賀田:"WWW検索のための情報収集技術の開発," 情報処理学会第54回全国大会,5L-02,1997.

[4]森本,間瀬,辻,絹川:"新聞記事自動分類システム構築の検討と評価," 情報処理学会第53回全国大会,3T-9,1996.

[5]宮井,徳永:"WWW情報検索システムにおける検索支援技術の開発," 情報処理学会第54回全国大会,5L-01,1997.