

# インターネット情報探索システム TITAN と その可視化手法

3 P - 7

鷲崎 誠司 林 良彦 菊井 玄一郎

NTT 情報通信研究所

E-mail : suzaki@isl.ntt.jp, { hayashi, kikui } @nttnly.isl.ntt.jp

## 1 はじめに

World Wide Web(WWW) 上の情報空間の拡大に伴い、いわゆる情報資源の探査の問題が顕著化し、WWW 空間からの効率的な情報探索手法に関して議論が進められている[1]。

情報量が増大し、更に内容が多様化している情報空間の効率的なアクセスのためには、従来の情報探索手法はそのまま適用できない。本稿では、単なる情報空間の探索だけではなく情報空間の知的アクセス手法の実現を目指したインターネット情報探索システム TITAN (Total Information Traverse AgeNt [2]) に関して述べる。

## 2 インターネットロボットの働き

WWW 空間は、世界中に広がる情報空間である。世界中から 1,600 万もの WWW ページを収集したという報告もあり、広大な情報空間に対して効率良く検索を行なうためには、予め情報資源を巡回し、内容を圧縮情報として手元に蓄積しておく、その情報に対して検索を行なえば、近似的に WWW 上の情報探索が可能となる。このしくみを持つシステムは、一般にインターネットロボット[3](以下ロボット)と呼ばれる。

ロボットの情報収集の方略は単純で、まずある起点となる URL(Uniform Resource Locator) のリストをロボットに与えると、それらのページからハイパリンク情報を取得し、URL を再帰的に訪問する。理論的な終了条件は、全ての有効なリンクを辿り終えた時と考える。この方法では、WWW 上に存在する全ての情報の収集は不可能であるが、全体から見ればかなり大きな部分集合を得られる。現在、TITAN では世界中から約 25 万件分のデータを収集している。次に、こうして得られた情報から、タイトル、ヘッドライン、テキスト段落、ハイパリンク注釈情報を分析する[4]。

## 3 多言語検索支援機能

多くの言語が存在する WWW 情報空間を、母国語でアクセスしたいという要求が多い。現状の WWW 上の情報は、ほとんどが英語で記述されているため、本システムでは、まず日本語の質問文を入力し、それを英語に変換することによる多言語検索支援を実現した。

検索式は、入力された語句(単語、句、文章も許す)を Juman を用いて形態素解析し、助詞など検索の際に悪影響を与える単語を除去した上で、原則的には最長一致による辞書引きで英語に変換する。変換時に多義が出る場合があるが、これは WWW から収集した情報からの頻度情報を用いて解消する。また、多くの言語と様々な文字コードが存在する WWW

情報の検索を裏で支える機能として、WWW 情報の文字コードや言語を自動的に識別する機能を備えている[5]。

## 4 ブラウジング支援付きインターフェイス

### 4.1 一覧できる GUI

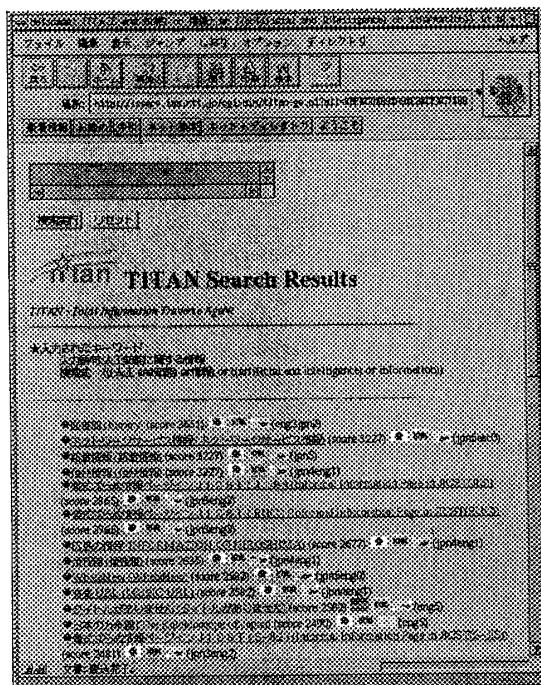


図 1 : TITAN の検索画面

従来の情報探索システムにおける検索結果は、検索語句にマッチする度合を得点化した上で整列表示していた。しかし、得点だけに基づく表示では、結果を一覧した際にどの行が本当に必要なものであるかがわかりづらいという問題があった。

本システムでは、情報発信国、情報のタイプ等を行の終りにアイコン表示することで、各々の情報の特徴を一覧すればわかるようになった。図 1 は TITAN の検索結果の表示画面、図 2 は検索結果の 1 行分を表している。図 2 における各行は WWW ページの情報を示しており、その表示形式を以下の通りである。

• インターネットカフェ (internet cafe) (score 2009)	eng(jpn)
1	2
3	4 5 6 7

図 2 : 検索結果の表示例

1 は日本語タイトル、2 は英語タイトル、3 は、検索エンジンによる得点化の結果である。4 では、情報発信国を国旗で表した。5 の情報形式は、HTML(Hyper Text Markup Language)、画像ファイルなどのファイル形式である。6 の情報の型は、情報収集の際に予め決定した尺度で、その情報がそれ自身で情報を持つ(情報型)ものか、それとも他の情報へのリンクが多いもの(案内型)かをアイコンで表している。全体で 2 つの情報を表す枠内グラフになっており、上部の青い部分は情報型の割合、下部の赤い部分は案内型の割合を表す。7 の記述言語は、情報がどの言語で記述されているかを自動的に判別し、複数言語で記述されている場合はその割合を表示する。

#### 4.2 情報の関連付け

WWW 空間は、いわばリンクで相互接続される広大なハイパーテキスト空間と考えられる。ハイパーテキストが持つ問題点に関しては、従来から指摘されているが、その中でも、空間内で自分の居場所がわからなくなる迷子問題や、一度に認識すべき事項が多過ぎる多選択肢などの問題は、WWW ブラウザが持つ機能だけでは解決できない。例えば、現在の WWW 上の情報探索システムは、キーワードを入力してそれに見合う情報を独自の得点化モジュールによる結果に従って出力するが、その結果の中には相互に関連が強いものも含まれている可能性もあり、そのような情報は利用者に通知すべきである。本システムは、ブラウザの情報探索結果を視覚的に補助する可視化部分を備えている。この部分は、Tcl/Tk+ndbm パッケージを用いて記述しており、WWW ブラウザである Mosaic、Netscape から呼び出され、各々 CCI(Common Client Interface)、NCAPI(Netscape Client API) を利用して相互通信を行なう一体性を高めている。

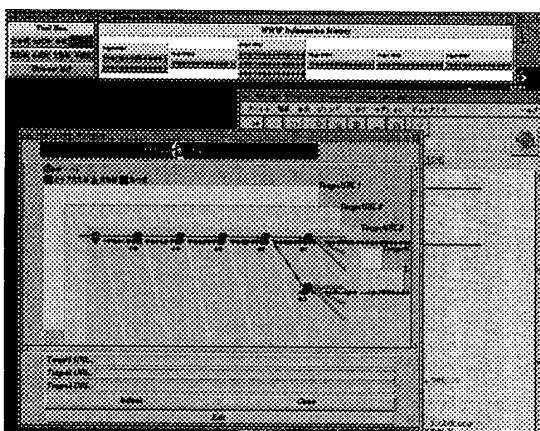


図 3 : 情報可視化の画面例

図 3 は、探索結果として導出された URL の配下にあるリンク情報、利用者の行動履歴の表示例を示す。URL の配下にあるリンク情報とは、情報探索結果として表示されている各行が、その配下にどのようなページ(URL)を持つかを示している。これは、ロボットにより予め収集しておいたページ情報からリンク情報を抽出しておき、各 URL がどの URL を含むか(正リンク情報)、そしてどの URL から指されているか(逆リンク情報)をデータベースに蓄積する。この情報を基にページ間の相互関係を表示することで実現した。また、回覧中に空間内で迷子にならないため、各 URL を直接操作しながら WWW ブラウザと連動して回覧していく際に、各々の URL 每にどの順にどの URL を回覧したかを履歴情報として表示し

た。本システムでは、現在までに正、逆合わせて約 180 万件のリンク情報をデータベース化している。

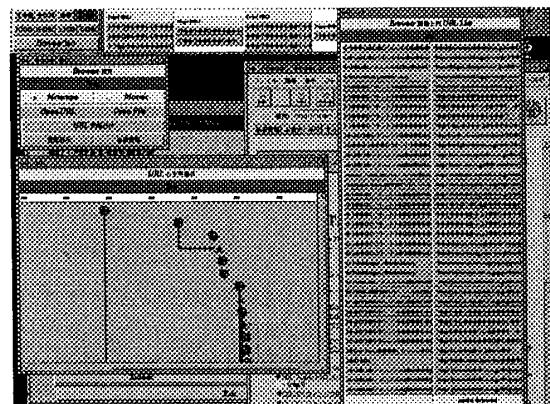


図 4 : 情報間の相互関連付けの例

図 4 は、各 URL 間の相互関係の表示例である。情報を表す URL を元に、そのパスの情報から、同じサーバー上にあるなどの情報をを利用して関連付けを行い、その結果を表示する。URL の標準形は以下のようなものである。

`scheme://host.domain[:port]/path[#anchor][?key]`

ここで `scheme` とは使用するプロトコルを表しており、`http`、`ftp` などがある。続く `host`、`domain` でアクセスすべきサーバを指定し、`path` でそのサーバ上でのデータの位置が示される。また、同じ URL の中においてマークされたポイントへのタグは `anchor` で、CGI スクリプトなどへの引数などは `key` で表すことができる。同じサーバー上にあるかどうかは、`host.domain` の部分を参照すればわかり、更に `path` 情報からサーバー上の上下関係がわかる。これを利用して相互関係を表示することで、利用者に対して情報を与えている。

#### 5 まとめ

本稿では、日本語を用いた WWW 上の情報探索システム TITAN の概要と、その可視化手法に関して説明した。

#### 参考文献

- [1] <http://rodem.slab.ntt.jp:8080/w3conf-bof/>
- [2] <http://isserv.tas.ntt.jp/chisho/titan.html>
- [3] <http://info.webcrawler.com/mak/projects/robots/active.html>
- [4] 林他: WWW 情報空間における Resource Discovery と Navigation 支援、電子情報信学会、人工知能と知識処理研究会、AI-95-31, 1995  
(<http://isserv.tas.ntt.jp/chisho/paper/9511Hayashi-SIGAI.ps.Z>)
- [5] 菊井他: インターネット情報ナビゲーションにおける多言語機能、情報処理学会、自然言語処理の応用に関するシンポジウム、1995  
(<http://isserv.tas.ntt.jp/chisho/paper/9511KikuiSIG-NL.ps.Z>)