解　説　素朴な疑問

WWWの検索サービスは何をしているのですか？

The Mechanism of WWW Search Services by Kent TAMURA (Tokyo Research Laboratory, IBM Japan, Ltd.).

田　村　健　人[1]

1 日本アイ・ビー・エム（株）東京基礎研究所

おそらくこの記事を目にされている方の多くが，WWWを日常的に利用しているかと思います．いつもみているページをみたり人から聞いたURLをみに行ったりするのではなく，「ある何かの情報が欲しい」というときにはWWW検索サービスを利用することになります．検索サービスを使いこなすことで欲しい情報を的確にみつけることができます．WWWにおいて重要なサービスである検索サービスは，どのような仕組みなのでしょうか．

検索サービスの仕組みの前に，WWWの仕組みについて理解しておきましょう．ここではHTTPについてのみ説明します．ブラウザが，あるURL http://www.foo.bar/baz.html を表示するには

(1) WWWサーバwww.foo.barに接続
(2)「/baz.html をください」と要求を出す
(3) 返ってくるデータを受けとって，表示

という手順を行います．ここで重要なのは，この手順はWWWサーバとブラウザの1対1の接続ですむことです．この手順の中では，「複数のWWWサーバからデータを探し出す」ということは考慮されていないのです．

検索を行うことを考えると，実際に検索を行うホスト上に検索対象のデータがなければなりません．検索要求を受けてから各WWWサーバにデータをとりに行くのでは遅すぎます．また，WWWサーバ内だけの検索も，標準的な手順が決められているわけではありません．

1. WWW検索サービスの種類

検索サービスは，「何を検索対象とするのか」「検索対象をどうやって調達するのか」によって大きく2種類に分類できます．

- イエローページ型(ディレクトリ型)

Yahoo!に代表されるタイプです．これはジャンルごとに分類されたリンク集であり，そのリンク集を対象に検索をすることができます．

WWW上で公開されているすべてのウェブページが検索対象になるわけではありません．あるウェブページが検索対象になるためには，誰かが能動的に登録しなければなりません．ウェブページの内容について検索を行うのではなく，登録されたタイトル・概要文などが検索対象です．

- WWWロボット型

「WWWロボット」と呼ばれるプログラムを用いてウェブページを網羅的に収集しておき，それを検索対象とするタイプです．AltaVista，gooなどが有名です．

ウェブページを収集するので，ウェブページの全文を検索対象にすることができます．ただし検索サービスによって，全文を洩れなく検索対象とする・先頭の数語のみ対象とする・タイトルと見出しのみ対象とする，などの違いがあります．

WWWロボット型は，イエローページ型に比べて桁違いの量のデータを扱います．ネットワーク・ディスク・計算資源のどれもWWWロボット型の方が多く必要でしょう．

イエローページ型は，人が選んだウェブページのみが登録されています．WWWロボット型はプログラムにより機械的に収集されたウェブページが登録されています．一般的には，イエローページ型サービスの方が検索結果の中に「外れ」が含まれる可能性は低くなります．

検索語を複数の検索サービスで検索し，その結果を統合して表示する「メタサーチエンジン」と呼ばれるものもあります．

2. WWWロボット

WWWロボットは，ウェブページを自動的に収集するソフトウェアです．

(1)あるウェブページのURLを未取得リストに入れる
(2)未取得リストからURLを1つ取り出し，そのウェブページを転送する
(3)転送したウェブページの中のリンク情報を抽出し，リンク先のURLを未取得リストに入れる
(4)(2)から繰り返し

という動作をします．これだけの単純な動作で，世界中のウェブページのほとんどを収集することができます．

フリーソフトとして配布されているWWWロボットもありますが，ほとんどの検索サービスは独自にWWWロボットを開発しています．

WWWロボット型の検索サービスにおいて，あるウェブページが検索できるかどうかはそれぞれのWWWロボットの判断によります．基本的には，すべてのウェブページを収集しようとするWWWロボットが多いのですが，「リンクされている数が多いウェブページを優先する」「サーバのトップから辿れるウェブページしか収集しない」などの方針がある場合もあります．

WWWロボットによる収集は時間がかかり，リンクされていないウェブページには辿り着けませんので，新しく作ったウェブページが検索できるようになるまでは数週間から数カ月の時間がかかります．そのため，WWWロボットが優先的にくるようにURLを登録することができる検索サービスもあります．

また逆に，勝手に検索サービスに登録してほしくないウェブページのために，WWWロボットの行動をウェブページ作成者が抑制する方法が決められています．

3. 検索サービスの動向

とくにWWWロボット型の検索サービスにおいては，「いかに高速に検索できるか」「いかに多くのウェブページを対象としているか」が具体的でわかりやすい評価基準でした．現在では次のようなことが課題となっています．

• ユーザにうまい検索語を入力してもらうためにはどうしたらよいか

単純に文字列のマッチングを行うのではなく，検索語の「意味」まで考慮するアプローチもあります．

• あまりにも多い検索結果をどうするのか

検索結果をどうみせればよいのか．簡単に絞り込み検索をするにはどうするか．

RCAAU Mo-n-do-uとODINは，検索結果と一緒に「検索された語と同時に出現することが多い語」による絞り込み検索候補を提示してくれます．AltaVistaの“Refine”では語どうしの関係をグラフ化して示します．

• 増え続けるデータに対してどう対処するのか

分散して検索できないか．WWWロボットですべて集めようと努力するのか，取捨選択するとしたらその方法はどうするか．データが増えてもイエローページの質を保てるか．

参考文献

1) Yahoo!, http://www.yahoo.com/
http://www.yahoo.co.jp/
2) AltaVista, http://altavista.digital.com/
3) goo, http://www.goo.ne.jp/
4) RCAAU Mo-n-do-u, http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/
5) ODIN, http://kichijiro.c.u-tokyo.ac.jp/odin/
6) Search Engines in Japan, http://www.ingrid.org/w3conf-bof/search.html

(平成9年7月30日受付)

田村 健人(正会員)
1972年生．1995年早稲田大学理工学部情報学科卒業．1997年同大学院理工学研究科情報科学専攻修士課程修了．同年日本アイ・ビー・エム(株)東京基礎研究所に入所．e-mail: kent@trl.ibm.co.jp