

WWWを用いた情報検索

WWW(インターネット)は情報の宝庫

インターネットでは、Web の仕組みを利用して様々な情報が提供されています。今日では、世界中に多数の Web サーバが立ち上げられており、Web サーバの中には一台でありながら膨大な量の情報を公開しているものもあります。

サーバ数に関する統計情報: <http://www.netcraft.com/survey/>
URLの数(ページ数等に関連)は、Googleによれば16億以上あるそうです。

このように情報がいわば氾濫しているような状況の中では、自分の必要としている情報を時間をかけずに的確に見つけ出すことは容易ではありません。また、情報を提供している Web サーバの存在自体を知らなければ、情報を探し出すことさえ困難です。

サーチエンジンの登場

そこで、インターネットに溢れる膨大な情報の中から自分の必要とする情報を容易に見つけ出すためのサービス、すなわち検索サービスがインターネット上で次第に提供されはじめました。このようなサービスを提供するサーバは「サーチエンジン」とも呼ばれていますが、サーチエンジンは定期的に世界中の Web サーバをリンクを辿りながら巡回し、各ページに掲載されている情報を自分の持つデータベースに登録していきます(Web のページを持つ情報提供者からのリクエストで登録される場合もあります)。

サーチエンジンがユーザから検索の要求を受けたときは、自分の持つデータベースの中で関連性の高そうなページをいくつか拾いだし、そのページの URL をユーザに返します。もちろん、機械による自動処理で得られる結果なので、得られた URL が必ずしも的確な情報原を指し示しているわけではありません。ユーザは提示された URL のリストから自分の必要としている情報に近いものを探していくことになりますが、ユーザが全て手作業で情報を探し当てる場合と比較すると、作業効率是非常に高くなっています。

検索は、基本的に、ページ内に含まれる文字列のキーワードマッチによって行われますが、キーワードの位置や、他のURLとのリンク関係、さらに、同じキーワードを用いて検索したユーザがどのようなページを優先的に参照しているか等といった、様々な付加的情報を加味した高度な検索が行われるようになってきています。

サーチエンジンの使いわけ

たいていのブラウザには検索機能が備わっていますが (IE: Internet Explorer でアドレスのかわりにキーワードを指定する等)、これは単に、ブラウザがインターネット上の特定のサーチエンジンに検索要求を転送するようになっているだけです。現在では、多数の検索サービスがインターネットで提供されていますが、それぞれ異なる特徴を出して差別化を図っていますから、調べようとしている情報の内容によってサーチエンジンをうまく使い分ける必要があるでしょう。また、一つのサーチエンジンで、インターネット上のすべての情報が網羅されているとは限らないので、あるサーチエンジンで必要とする情報が得られなかったからといって、インターネット上にそのような情報を提供しているページが存在していない、ということにはならないことに注意する必要があります。

古いURLに注意

なお、サーチエンジンを利用して得られた URL はすでに古くなっていて、対応する Web のページが無くなっていたりする場合があることにも注意する必要があります。サーチエンジン

における検索では、それぞれのサーチエンジンが持つデータベースが元となっているわけですが、このデータベースの内容の更新頻度がどれくらいであるかがポイントとなります。更新頻度が低ければ情報の新鮮度が低下してしまうわけですが、更新するためには世界中の Web サーバの全てのページを順にアクセスしていく必要があるため、更新頻度があまりにも高いと無駄な通信が多くなってしまいます。そうなると、各 Web サーバの負荷が不必要に高くなったり、本当に必要な通信が妨げられたりする可能性があるため、更新頻度を不必要に高くすることは得策ではありません（迷惑をかけてしまうことになります）。このようなことも念頭におきながら、サーチエンジンを上手に活用してください。（とはいいつつ、検索対象となる情報の新しさを売りにしたサーチエンジンもあります。）

サイト内検索

なお、一般の Web サーバの中には、その Web サーバ(サイト)が提供している情報に関して検索サービスを提供しているものも次第に増えてきています。自分の知りたい情報が、どの Web サーバにあるのかが特定できているような場合(例えば、ある会社の商品に関して知りたい場合など)は、一般的なサーチエンジンを利用する代わりにそういった Web サーバ付随の検索機能を利用する方法もあります。

(一般的なサーチエンジンでも、検索対象とするページを、特定の Web サーバに限定する機能を持っているものもあります。)

また、インターネット上には、Web のページを検索の対象とするサーチエンジンだけでなく、従来からインターネット以外を利用する方法で提供されていた図書文献等の検索が可能なサーバも増えてきているので、そのような検索サービスも活用すると良いでしょう。

サーチエンジンのリスト

たくさんあるので、サーチエンジンも自分で探してみましょう。ここでは、サーチエンジンを探すための参考ページをいくつか示しておきます。

- [検索デスク \(http://www.searchdesk.com/\)](http://www.searchdesk.com/)
- [Yahoo のサーチエンジンに関するページ](#)

なお、メタサーチとは、サーチエンジンに関する情報を集め、サーチエンジンを検索したり、複数のサーチエンジンを同時に利用して検索の効率を高くするためのものを指します。

検索の手法に関する分類

まず、検索の手法に関する基本的な分類を整理しておきましょう。

- カテゴリ検索(ディレクトリ型/主に手動登録)
 - 逆引カテゴリ
 - 総合索引みたいなもの
- キーワード検索(ロボット型/主に自動登録)
 - 前方一致/後方一致
 - AND/OR/NOT (条件の論理演算)
一般に、複数のキーワードを単に列挙すると AND 検索となる
 - 絞り込み検索
検索結果に対してさらに検索を行う
- URL逆リンク検索
指定した URL を参照しているページを探す。
- メタサーチ
複数の検索エンジンを同時に利用するフロントエンド

この他、検索できる内容での分類もあります。

書籍検索

書籍を販売しているところの中には、書籍の検索サービスを提供しているところがあります。
(書籍のインターネットショッピングサイトでは、検索機能が提供されています。)

- <http://www.books.or.jp/>
- <http://www.kinokuniya.co.jp/>
- <http://www.kyoto-bauc.or.jp/shop/>
- <http://www.yaesu-book.co.jp/>
- <http://www.amazon.co.jp/>
- http://www.trc.co.jp/trc-japa/search/trc_www.htm

蔵書検索 (OPAC: On-line Public Access Catalog)

図書館も、インターネットから蔵書検索ができるようになってきています。

- <http://www.ndl.go.jp/>
- <http://kensaku.libnet.kulib.kyoto-u.ac.jp/>
- <http://ss.cc.affrc.go.jp/ric/opac/opaclist.html>

シソーラス

インターネットからアクセスできる辞書もあります。またシソーラス(義語／広義語／狭義語／関連語等を調べるための辞書)もあります。

- <http://www.gengokk.co.jp/thesaurus/>
- <http://digilib.silkroad.net/thesaurus/>

地図

地図情報等もインターネットで提供されています(著作権等に注意すること。もちろん地図だけが問題になるわけではありませんが)。

- <http://www.mapion.co.jp/>
- <http://www.mapfan.com/>
- <http://www.jorudan.co.jp/>

whois

インターネットに関する情報は、JPNIC の whois で調べるのが基本ですが、他にも情報提供しているところがあるようです。

- http://www.nic.ad.jp/cgi-bin/whois_gate
- http://www.mse.co.jp/ip_domain/index.shtml

記事検索

新聞記事の検索など。

- <http://www.mainichi.co.jp/>
- <http://www.sankei.co.jp/>
- <http://www.kyodo.co.jp/>
- <http://www.goo.ne.jp/>

- <http://www.infoseek.co.jp/>
- <http://www.watch.impress.co.jp/internet/>
- <http://www.goo.ne.jp/>

CD-ROM検索

最近では、CD-ROMでも様々な情報が提供されています。場合によっては、CD-ROMを検索して必要な情報を入手することも必要となるでしょう。

京都大学の附属図書館では、CD-ROMサーバを用いた検索サービスも提供しています(学内からの利用のみ)。<http://www.kulib.kyoto-u.ac.jp/guide/jpn/current/g090j.html#G10P3>

ポータルサイト

ブラウザを起動したときに最初に表示されるページとして利用してもらうことを目的として運営されているサーバのことです(ポータル=入口)。ニュース提供や検索などの参照頻度の高い内容を中心にまとめられています。

例えば、次のURLを見るといくつか紹介されています。

- <http://dir.biglobe.ne.jp/dir/182186/178581/180227/180248/>
- http://dir.yahoo.co.jp/Business_and_Economy/Shopping_and_Services/Internet_Services/Portals/

FORMとCGI

WWWにおいて、ユーザからの入力を受けて何らかの処理を行い、処理結果を表示するようなページは、HTMLのFORM書式と、WWWサーバのCGI(Common Gateway Interface)機能を用いて作られています。

今回紹介した検索のページも、検索のためのキーワードを入力するところがありますが、この部分も一般にFORMやCGIによって作られています。実際に自分で同様のページを作ろうとする場合には、WWWサーバがCGI機能を提供しているかどうか確認しておくことが必要です。

本日の課題

- 地図で自分の家を探し、見つかった地図をパソコンに保存する
- いくつかのサーチエンジンを利用してみる
 - 他にどのようなサーチエンジンがあるか調べる
 - それぞれのサーチエンジンの特徴を調べる

検索項目の例：

- インターネットの発展に関する統計情報
- ネチケット
- プライバシー
- インターネットと著作権の問題
- インターネットを活用した学校教育
- 社会におけるインターネットの活用

レポート課題

インターネットのサーチエンジンを利用して、自分なりに興味をもっている事柄やサービスを

検索して見つける。

できるだけ、講義で紹介していない検索エンジンや方法を見つけて利用するようにすること。

レポートとして次の内容についてまとめ、電子メールで提出する。提出方法は、第2回のレポートと同様に、テキスト形式の添付ファイルによるものとし、レポートの内容はすべて添付ファイルに記述すること。

提出するレポートの内容はおおよそ以下のとおり。

- 回生、学生番号、氏名
- 検索等で見つけたページの URL
- そのページの概要説明
- なぜそのトピックについて検索を行おうと考えたのか
- 検索結果の中からはなぜそのページを選んだのか
- どのようにしてそのページを見つけたのか（手順等を詳細に）
（見つけるための自分なりの工夫を重点的に）
- jhg-report@media.kyoto-u.ac.jp へてにメールで提出
- 提出したものは、<http://quena.media.kyoto-u.ac.jp/jhg/3/>にて参照可能（提出確認）
- 締め切りは、6/19 20:00 とする