# 目指せ!ネットエスパー
# インターネット新検索術

インターネットの中に潜む膨大な情報たち。そこから欲しいものを見つけ出して活用するには、もはや1検索サイトだけでは役に立たない。ネットの海を自在に泳ぎ、必要な情報をすぐ取り出して活用する、そんな「ネットエスパー」に変身すればインターネットの利用価値は無限大になる。この連載で「ネットエスパー」に変身するスタートを切ろう!

二木麻里（アリアドネ運営）　Jump ariadne.ne.jp
Illust : Ebisu Yoshikazu

## 最終回　検索サイトを極める

検索サイトは現在、日本国内だけでも優に20を超す。サイトそれぞれの漠然とした違いは感じても、適当にキーワードを入れるとそれなりにヒットがあるため、ついいつも同じ検索サイトで同じ探し方をしてしまったりする。慣れない検索サイトのヒット結果には、なんだか違和感を持ったりするものだ。しかし構造的な違いや共通点がわかれば、さまざまな検索サイトを使い分けられるのでは？「インターネット新検索術」もいよいよ最終回。改めて探し物の出発点、検索サイトの中を探してみよう。

## 1　データベースを「透視」する

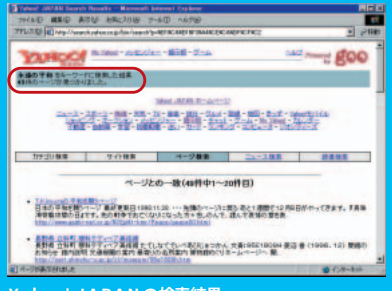
検索サイトというと、とかくサーチ機能の能力に目がいくが、検索結果に直結する部分だけに自然なことかもしれない。しかしどれほどサーチ機能が優れていても、サイトが持つデータベースに含まれていない情報は決してヒットできない。検索サイトのもっとも巨大な部分はデータベースである。

したがって検索サイトの第一の決め手は、データベースがどのように構成されているかにかかっている。ユーザーは、その検索サイトの構造、つまりデータの持ち方をできるだけ理解し、「透視」したい。

検索サイトのデータには、ロボットによってランダムに蓄えられたものと、人の手でカテゴリー分類されたディレクトリー構造のものがある。現在多くの検索サイトは、この対照的な2種類を何らかの形で併用しているわけだが、各サイトの特徴はこの併用の仕方にある。サイト内部で最初から複数のデータベースを構築している場合や、他の検索サイトと協力して相互補完関係を作っている場合も多いのだ。

たとえばディレクトリー系検索サイトである米国Yahoo! Jump01 は、大規模なロボット系サイトGoogle Jump02 の結果を取り入れて表示している。またLycos Jump03 は、当初ロボット系サイトとしてスタートしたが、のちにディレクトリー系に構造を変更し、FASTSearch Jump04 からの検索結果でデータを補っている。日本では、ほぼ完全なロボット型のgoo Jump05 と、ディレクトリー系の代表

gooの検索結果
Jump05 www.goo.ne.jp



Yahoo! JAPANの検索結果
Jump06 www.yahoo.co.jp

ロボット：ウェブ上を巡回しページのデータを拾い集めるプログラム。ロボットが自動的に集めたページを対象にして検索を行うサイトを「ロボット型の検索サイト」と呼ぶ。これに対して、スタッフが手作業で登録したページを検索するサイトは「ディレクトリー型の検索サイト」と呼ばれる。

Yahoo! JAPAN Jump 06 との連携が有名だろう。

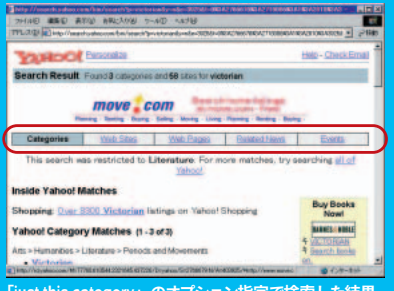
別サイトで探したのにまったく同じ結果が出たり、1つのサイト内で探しても違う結果になったりするのは、まずこのようなデータベース同士の関係による。これにトラフィックの状況など複数の要素が加わりはするが、検索サイトに神秘はない。

Jump 01 www.yahoo.com
Jump 02 www.google.com
Jump 03 www.lycos.com
Jump 04 www.ussc.alltheweb.com

## 2 カテゴリー内を検索する

わかりやすい例で見てみよう。たとえば米国Yahoo!の場合、どこまでがYahoo!で、どこからがGoogleなのだろうか。まずご存じのように、Yahoo!ではウェブページの性質を詳細に区分している。ツリー状の階層構造が非常に深くまで枝分かれするため、たどるのに忍耐がいるといわれがちだ。だが、カテゴリーの各階層ごとに1つずつ検索サイトがあると考えるとむしろ使いやすい。この小データベースの内部にだけ検索をかける機能も用意されている。

Yahoo!でディレクトリーを下りると、そのたびに「just this category」（このカテゴリーの中だけを探す）というオプションが検索窓の脇に掲示される。これは多くのサイトに見られる基本的なオプションで、階層を下るにつれて次第に小さくなるデータ空間を想像するとわかりやすい。ユーザーは途中まで階層を降りてからキーワード検索を行うことで、誤った領域のデータベースからあがるノイズヒットを減らすことができる。そして、データベースが小さいぶん検索スピードも速い。このオプションは、日本語圏のディレクトリー系サイトではヒットする資料そのものが少なすぎてさびしい場合もあるけれど、巨大なウェブデータを持つ英語圏などでは本領を発揮する。

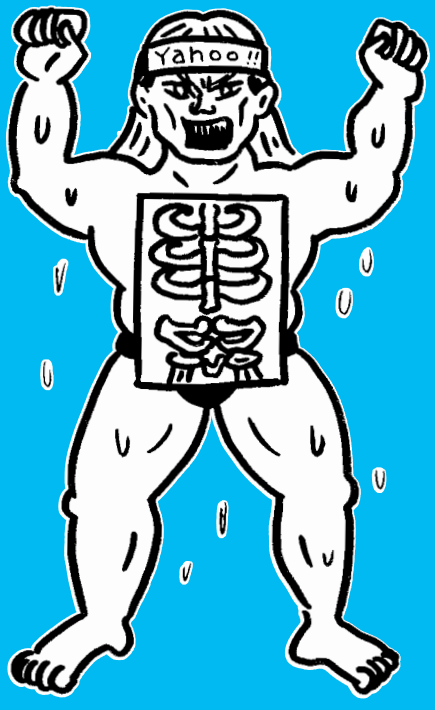


「just this category」のオプション指定で検索した結果。ここから別のオプションにワンクリックでたどれる。

たとえば米国Yahoo!でLiteratureのカテゴリーを開いて、「victorian」とキーワードを入れる。Yahoo!ではアルファベットの大文字と小文字は識別しないのでVictorianでも同じだが、ともあれ「all of Yahoo!」（Yahoo!サイトの全体を探す）というデフォルト設定による検索では5つの別カテゴリーが該当し、ビクトリア朝時代の「文学」以外に「歴史」、「地域研究」などのカテゴリーがリストアップされた。また同時に、サイト単位の検索結果が表示され、1916件がヒットしている。これらは基本的に、人の手によってYahoo!に登録されたサイトだ。

元のページに戻って、同じ検索を今度は「just this category」のオプション指定で行うと、カテゴリーヒット数は3、サイトヒット数は58。1916件対58件というデータベースのサイズ差が、質の高いヒットであることを期待させる。さらにこのときページの上部に、別のオプションも出ているのに注意したい。「Web Sites」（登録されたサイトの検索）「Web Pages」（ロボット検索）「Related News」（関連ニュース）などはここからワンクリックでたどれる。「Web Pages」の結果は22万7000件。これがGoogleを基盤にしたロボット検索の結果だ。ページ上部にはきちんとPowered by Googleと掲示されており、全体的にロジックの透明度は高い。

検索サイトは、このように多数のデータベースの束でできている。たとえば米国Yahoo!には人物サーチ機能「People Search」がある。これは固有名詞を入れるとその住所や電話番号、メールアドレスを探す機能で、カスタマイズすれば自分のアドレスブックとして使える独立したデータベースである。一方、Yahoo!の一般データベースで固有名詞をひくと、カテゴリーにヒットしなければ通常Googleから拾ってくる。




基本的にどの検索サイトもこうした「穴のない検索」を目指している。だがデータベースそのものに大差はなくても、実際にコンパクトなナビゲーションと充実した検索結果を生み出せるかどうかは、データベース同士の重層構成やサイト設計の巧みさ、そしてウェブサーフスタッフの陣容による。情報の評価能力と編集能力はヒューマンスキルがもたらす決定的な部分だ。Yahoo!はオフィシャルサイトが常に出てくるという安定性が高く評価されているが、それも人の手による分類作業が正確に行われていることによる。

ディレクトリー：コンピュータの世界で、ディスクなどにあるファイルを管理するための仕組み。ツリー構造をした元ディレクトリーによってファイルが階層的に管理されている。検索サイトにおいてサイトをカテゴリーごとに分類、整理する方式を、ロボット収集型のサイトに対してディレクトリー形式と呼ぶことも多い。

# 3 キーワード認識の違いをつかむ

データベースはいわば膨大な言語の塊だ。ことに和文の場合、表記上は単語同士の切れ目がない。そこから望む言葉だけを小さく切り出してくるプロセスこそが「検索」である。無駄なく正確にヒットするように、検索サイトのプログラム内部ではあらかじめ一定の単位で単語をインデックス化している。この単語の切り分け方が検索サイトごとに異なるため、私たちはよく混乱する。これは人によって野菜の切り方が違うようなものだ。また、句読点やスペースはどのように認識されるのか、異表記への読み替えは行われるのか、「と」や「の」などの接続詞や助詞は意味を持つのか、また「+」や「""」などの検索用記号は何を使えばいいのかなどが問題になる。

キーワード認識にはおおまかに区分けすると、文章をまるごと認識する「フレーズ対応」（テクスト認識）と単語を切り離して認識する「ワードブロック対応」とがある。Yahoo!やgoo、 Googleは前者、Lycos、exciteは基本的に後者の比重が高い。カントの『永遠平和のために』を何通りか検索してみた。

Yahoo!はgooを取り入れているので、基本的に同じ検索結果を出している。助詞なし、助詞あり、文章、および中黒（・）入りのすべてに反応しており、どの要素も検索条件として認識している。スペースは「キーワードすべてを含む」と解釈される。また、単語の部分入力にも対応している。gooの場合、デフォルトの条件指定は「すべての語を含む」だが、事実上「フレーズ指定」と同じ結果になることも多い。

Googleは中黒を落として表示してくるが、あらかじめ中黒を落とした表記「エマヌエルカント」では0件。つまり、中黒を認識している。部分検索も行えるが、再現性は必ずしも高くない。

Lycos Jump01 とexcite Jump02 はデフォルト設定ではワードブロック対応のため助詞は認識されないが、条件指定のページに移ってフレーズ指定を行うと、初めて認識される。「永遠平和」より「永遠平和 カント」のほうが検索結果が多くなるのは、スペースが「キーワードのどれかを含む」として認識されるためである。

Jump01 www.lycos.co.jp
Jump02 www.excite.co.jp

単位：件

| | Yahoo! Japan | goo | Google（日本語指定） | Lycos Japan | excite Japan |
|---|---|---|---|---|---|
| 永遠の平和 | 49 | 50 | 722 | 457 | 24万2703 |
| 永遠平和 | 15 | 14 | 143 | 8316 | 24万2703（フレーズ指定では15） |
| 永遠平和 カント | 10 | 10 | 108 | 142 | 24万6478（フレーズ指定359） |
| 永遠平和のために | 9 | 9 | 97 | 4036 | 24万2703 |
| エマヌエル・カント | 2 | 2 | 10 | 10 | 24万6775 |
| 永遠平和のために エマヌエル・カント | 0 | 0 | 0 | 0（フレーズ指定0） | 4123 |

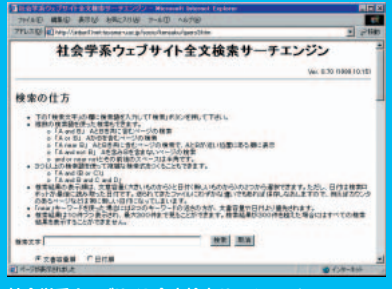
# 4 専門検索サイトを「透視」する

検索サイトにおいてデータベースが身体であるとするなら、サーチ機能は頭脳、巡回ロボットは心臓から送り出される血液かもしれない。データという身体細胞を洗って、新鮮なものに保つ。こうした検索サイトとしての基本的な仕組みは、汎用検索サイトでも専門領域の検索サイトでもそう変わらない。だが最大の違いは、やはりデータベースの特性だ。

この連載でも繰り返し取り上げてきたが、専門領域に特化された検索サイトが究極に目指すものは何だろうか。そしてユーザー側はそこにどういう検索結果を期待できるのだろうか。

1つの例として、「社会学系ウェブサイト全文検索サーチエンジン」 Jump の構造を見てみよう。個人が独力で制作している検索サイトだ。こうした良質のウェブサイトはほかにもあるが、ここは巡回プログラム部分までほぼ自作という希有なサイトである。巡回先のウェブサイトはすべて公開されており、いわばガラス張りの透明性を備えている。

そして、このサイトのタイトルはなぜ「全文検索」なのか。そこからはこのサイトが「資料を読むこと」、つまり文献性の高いオンラインリソースを探すことを念頭に置いていることがうかがわれる。インターネット上にある社会学系のサイトという検索範囲を前提に論文や報告書などを探すといっているのだ。解説によれば巡回対象は約5万2000ページ。インデックス作成プログラムの種類や読み込み対象エクステンション、また新ファイルのリストなどが端然と掲示されていて、自分が行った検索がどこからどうデータを引き出してきているのかを快く「透視」できる。




社会学系ウェブサイト全文検索サーチエンジン
Jump jinbun1.hmt.toyama-u.ac.jp/socio/kensaku/query3.htm

デフォルト：プログラムにあらかじめ設定してある標準値。コンピュータやソフトウェアなどの場合は、出荷時に設定してある値。

「出版史」を検索すると、ヒットは1件。博士論文がすぐ呼び出されてきた。きちんとアブストラクト（摘要）がついていて概要を読める。論文のタイトルは「漢語の影響下におけるモンゴル語近代語彙の形成―中国領内のモンゴル語定期刊行物発達史に沿って―」。正確なヒットだ。もう1つ「出版 and 戦後 and ハーバーマス」で検索すると44件。ウェブサイトの最上部にリンクせず、そのまますぐに本文を読んでいくことができる。

巨大で複雑な汎用検索サイトは、「データベース内のノイズをいかに拾わないか」を考えて作られている。だが専門検索サイトは、最初から「データベース内にどうやってノイズを入れないか」と考えてくる。おそらく、求める理想は“ノイズ・ゼロ”である。

ぜひ一度、こういう手作りの専門検索サイトを使ってみることをおすすめしたい。ときに複雑怪奇にすら思える汎用検索サイトの仕組みも理解しやすくなることだろう。そう、検索サイトってこういうものだったんだと。いっそ自分で検索サイトを作れば、それこそ究極の「新」検索術かもしれないけれど……。

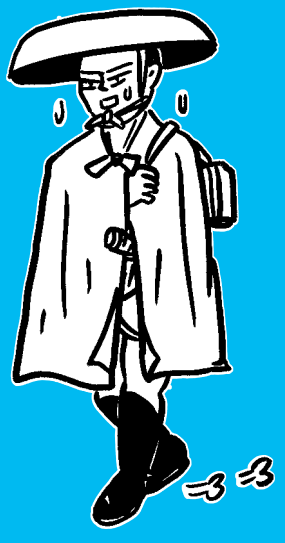
## 検索をより好きになるサイト

探し物が苦手な人にも、検索なら任せて、という人にもおすすめ。ネットで得られる日々のニュースと、そうした情報を探すヒントを一緒に読める場所が「インターネットと情報検索／きになるWEB」Jump 01。関裕司氏による、ユニークで親しみやすいサイトだ。毎日更新されるニュースクリップに加え、「情報検索ノウハウ集」のページには「探したいサイトにたどりつくための七つの教え」など、検索の基本がわかりやすく紹介されている。

また、実際に検索サイトを利用して探すという切り口では、浅井勇夫氏の「検索デスク」Jump 02 が有名だろう。各種の検索サイトがぎっしり掲げられ、パラレルサーチとして機能するほか、関連情報や多数のニュースを探せる。すべてをぎゅっと1ページに収めた簡潔な表形式が印象的。「きになるWEB」も「検索デスク」も、目的と手段の両方を手際よく1つの情報箱にまとめていて、きびきびした魅力が伝わってくる。

Jump 02 www.searchdesk.com

インターネットと情報検索／きになるWEB
Jump 01 www.shikencho.com



## 5 今月のポータルキット

調べ物の始まりは、汎用検索サイトだ。そこから出発する「探し物の旅」に原則があるとすれば、情報が集約された「島」をできるだけ早く確保することかもしれない。それはしばしば専門領域のゲートサイトである。たくさんのサイトをザッピングしながら次々とさまよう前に、しっかり中心になるウェブサイトをまずは1つ、あせらずていねいに調べるほうがずっとよい。それから、そこを拠点にウェブサーフに出る。情報の「島」はブラウザーでキープしておき、迷ったらいつもここに戻ろう。またウェブサイトの性質は見分けにくいが、URLでも一定の見当はつく。きちんとURLを読むくせをつけると、島から外へ出るときも無駄なジャンプを減らせる。

こうした原則は、これまでの12回の連載でさまざまな形で取り上げてきた。終了後もすべての回をオンラインで読めるよう、インターネットマガジンのサイトJump に掲載したので、あらためて参考にしていただきたい。

Jump internet.impress.co.jp/esper/

| | |
|---|---|
| 出発 | 汎用検索サイトでコンパクトなサーチを試みる |
| 初級 | その分野の中心となるウェブサイトを「島」として確保 |
| 中級 | 専門検索や個人のウェブサイトを比較利用 |
| 上級 | オンライン辞書・事典を参照しつつウェブサイトを歩く |
| 原則 | ウェブサイト評価やウェブサーフで迷ったら、島に戻る |

**二木麻里（ふたき まり）**
上智大学外国語学部卒。翻訳家。社会・人文科学系の国内外資料を案内した総合サイトARIADNEを運営。著書に『思考のためのインターネット』（筑摩書房ちくま新書）など。

汎用検索サイト：ワールドワイドウェブ全体の中から、あらゆるジャンルの物事を探し出す目的で作られた検索サイト。代表例として日本では「Infoseek」Jump www.infoseek.co.jpや「goo」Jump www.goo.ne.jpがある。

# ［インターネットマガジン　バックナンバーアーカイブ］ ご利用上の注意

このPDFファイルは、株式会社インプレスR&D（株式会社インプレスから分割）が1994年～2006年まで発行した月刊誌『インターネットマガジン』の誌面をPDF化し、「インターネットマガジン　バックナンバーアーカイブ」として以下のウェブサイト「All-in-One INTERNET magazine 2.0」で公開しているものです。

http://i.impressRD.jp/bn

このファイルをご利用いただくにあたり、下記の注意事項を必ずお読みください。

- 記載されている内容（技術解説、URL、団体・企業名、商品名、価格、プレゼント募集、アンケートなど）は発行当時のものです。
- 収録されている内容は著作権法上の保護を受けています。著作権はそれぞれの記事の著作者（執筆者、写真の撮影者、イラストの作成者、編集部など）が保持しています。
- 著作者から許諾が得られなかった著作物は収録されていない場合があります。
- このファイルやその内容を改変したり、商用を目的として再利用することはできません。あくまで個人や企業の非商用利用での閲覧、複製、送信に限られます。
- 収録されている内容を何らかの媒体に引用としてご利用する際は、出典として媒体名および月号、該当ページ番号、発行元（株式会社インプレスR&D）、コピーライトなどの情報をご明記ください。
- オリジナルの雑誌の発行時点では、株式会社インプレス R&D（当時は株式会社インプレス）と著作権者は内容が正確なものであるように最大限に努めましたが、すべての情報が完全に正確であることは保証できません。このファイルの内容に起因する直接的および間接的な損害に対して、一切の責任を負いません。お客様個人の責任においてご利用ください。

このファイルに関するお問い合わせ先
株式会社インプレスR&D
All-in-One INTERNET magazine 編集部
im-info@impress.co.jp