「Googleはヒットする」の秘密も解明

検索エンジン解体新書

ふだんなにげなく使っている「検索エンジン」。検索エンジンは「ディレクトリー型」と「ロボット型」に分類されることが多いが、今回は現在の検索エンジンの主流である「ロボット型検索エンジン」の仕組みを解説する。仕組みがわかれば、ロボット型検索エンジンの得意な分野と苦手な分野がわかり、これまでよりスムーズに必要なコンテンツにたどり着けるようになるだろう。

集中企画

坂野幸臣＋原田昌紀＋高野 元＋鈴木基久＋編集部

人力と機械、そこが最大のポイント

検索エンジンの種類と仕組みを知る

検索エンジンには、人手で登録する「ディレクトリー型」と、コンピュータが自動で登録する「ロボット型」がある。ここでは、両者の違いと、ロボット型検索エンジンがどのようにサイトを登録して検索できるようにしているのかを解説しよう。 編集部

“人手vs機械”から“人手＆機械”へ

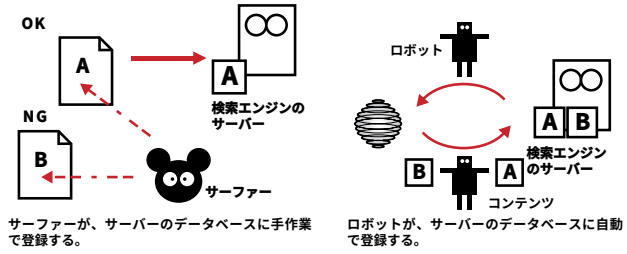
検索エンジンには、サイトの評価と登録・分類作業をサーファーと呼ばれる人が行う「ディレクトリー型」と、サイトの評価と登録をコンピュータが自動で行う「ロボット型」に大別されてきた（図❶参照）。

ディレクトリー型の検索エンジンでは、人が一度チェックしているので、良質なサイトがわかりやすく登録されるが、その半面で検索できる範囲が狭くなる。一方のロボット型は、検索できる範囲は広いが、関係の薄いサイトも大量にヒットするとされてきた。

ところが、現在のロボット型は旧来の弱点を人手でカバーしている。たとえば「首相官邸」のように「公式サイト」があるキーワードは、通常のデータベースとは別に人手で登録することで、公式サイトがヒットしやすいようにしている。

また、現在のロボット型はディレクトリーも持ち、該当するディレクトリーを検索結果に表示するようになっている。欠点をカバーしながらロボット型検索エンジンは進化しているのだ（図❷参照）。

図❶ ディレクトリー型とロボット型



「ディレクトリー型」は、サーファーと呼ばれる専門のスタッフがサイトを評価して手作業で登録する。一方、「ロボット型」はウェブを巡回して無作為にサイトを登録する。

サーファー：ディレクトリー型検索エンジンで、サイトを評価・分類するスタッフ。多くの場合、サイトの紹介文もサーファーが書いている。

ロボット型が検索できるまで

ロボット型は、大きく分けて4つの作業を経て、サイトを検索できるようになる（図❸参照）。

まず、「ロボット」と呼ばれるプログラムが、インターネット上にあるさまざまなサイトの情報を収集する。ロボットとは、訪れたサイトのリンク構造を抽出して自動的に「次に訪問するサイト」を決め、広範囲のサイトを収集するためのプログラムのことだ。この優劣が検索エンジンの更新頻度を決める。

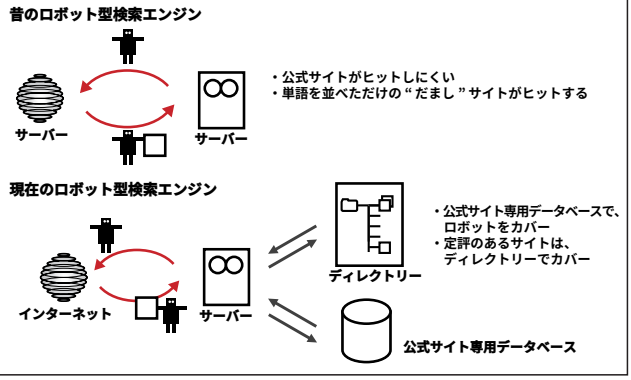
次に、ロボットが集めたサイトのデータを解析して、「インデックス」と呼ばれるデータベースに記録する。この作業は、検索エンジンによって解析方法やスピードや更新方法に差があり、各検索エンジンの特徴を左右する部分でもある。

それから、利用者が入力したキーワードに合うサイトをインデックスから探し、一定の法則で採点をする。この「採点方法」は各検索エンジンの心臓部で、常に新しい手法が開発されている。

最後に、検索結果をHTMLに出力し、利用者がブラウザーで見られるようにする。検索結果のデザインも検索エンジンの使い勝手を左右するが、この特集の趣旨から外れるので多くは解説しない。

それぞれの技術を次ページから詳しく説明しよう。なお、特に断りがなければ、本特集における文中の「検索エンジン」とは、ロボット型検索エンジンのことを指す。

図❷　ロボット型検索エンジンの進化



図❸　ロボット型検索エンジンの構造例

表示	表示エンジン		
ヒット判定	キーワードの形態素解析	隣接判定	スコア計算
インデックス	タグ除去	形態素解析	キーワード出現位置保存
収集・登録	ウェブサイト収集ロボット		

ロボットがウェブを巡回して、サイトを検索エンジンのデータベースに登録してから、検索結果として表示されるまでは、下の層から上の層への順番で処理される。

リンク構造：HTMLで「Aタグ」を使い、ハイパーリンクが張られているページの関係。多くのサイトはトップページから図形図のようにハイパーリンクが張られている。

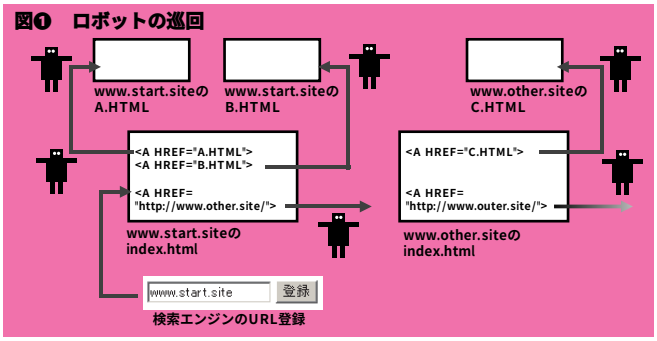
ロボットはどのようにサイトを見つけるか

ロボットの仕組みを理解しよう

現在の検索エンジンは、数千万から数億といわれるサイトを検索できるという。これは、ロボットというプログラムの働きによるものだ。ここでは、ロボット型の“ゆえん”たるロボットの特徴を解説する。

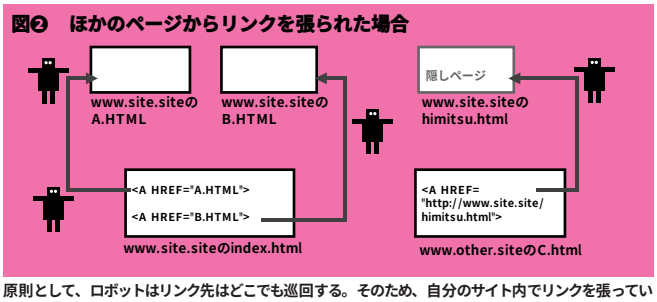
坂野幸臣

図❶　ロボットの巡回



ロボットは、HTMLのAタグを見つけるとURLを抽出して、次の巡回先として登録する。リンク先が違うサイトでも、自動的にトップページを見つけて巡回する。

図❷　ほかのページからリンクを張られた場合



原則として、ロボットはリンク先はどこでも巡回する。そのため、自分のサイト内でリンクを張っていなくても、別のサイトからリンクを張られれば、ロボットは巡回してしまう。

ロボットは自動でリンクをたどる

インターネット上のサイトから、自動で情報を収集する検索エンジンのプログラムを一般に「ロボット」と呼ぶ。最初に、ロボットは検索エンジンのサイトに用意してある「URL登録」ページで、ユーザーが登録したURLにアクセスして情報の収集を始める。ロボットがアクセスする行き先のURLからHTMLを取得して、そのHTMLの中でリンク情報として見つけたURLを抽出して、リンク先を次の巡回先として自動的に設定する。これを延々と繰り返すので、1つの登録URLから幅広いサイトを自動で巡回できる（図❶参照）。

集めたデータは「リンク」と「文書」を中心に解析される。「リンク」のURL情報はロボットが次に巡回する行き先となり、「文書」は検索用のデータとなる。一度ロボットが巡回したURLは、検索エンジンのサーバーにリスト化して保管され、サイト情報の更新のため、再度巡回するときにも使われる。

各サイトのHTMLをすべてデータベースに保存するのは現実的ではないため、収集した文書は適当に変換して保存されるが、「Google」のようにキャッシュとしてすべてを保存する検索エンジンも存在する。

「隠しページ」がヒットするのはなぜ

ロボットはリンクをたどって巡回するため、制作者が自分のサイト内でリンクを張っていない、いわば「隠しページ」もヒットすることがある。おもな原因は、掲示板やメーリングリストの過去ログやほかの個人サイトからリンクを張られているからだろう（図❷参照）。

ほかに、意外なところで「プロキシーサーバーのログ」が犯人になることもある。プロキシーサーバーのログがHTMLで保存されている場合、ロボットはプロキシーサーバーにあるURLをリンクとして抽出し、巡回する。

さらに、善意・悪意にかかわらず、他人に勝手に登録される場合もある。基本的に検索エンジンのURL登録は、管理者・制作者とは関係なしに誰でもURLを登録できるからだ。もちろん自動巡回による登録や、手動での登録を意図的に回避することは可能だ。知られたくないページは制作者が制御する必要がある。

Tips!

こんなページは検索できない

<META name="robots" content="noindex,nofollow">
❶　❷

❶	登録する	index	❷	リンクをたどる	follow
	登録しない	noindex		リンクをたどらない	nofollow

これらの処理は「紳士協定」に基づいて、検索エンジンでは検索されないようになっている。しかし、これだけでは検索エンジンに登録されないだけなので、特定の人だけが見られるようにするためにはCGIなどを使ったアクセス制御やパスワードによる認証が必要だ。

プロキシーサーバー：社内ネットワークなどからインターネットに接続するとき、内部のセキュリティーを守るために境界に置く代理サーバーのこと。また、ウェブアクセスにおいては、HTMLデータなどをキャッシュし、同様のリクエストに対して蓄積したデータを返すことによって、クライアントのウェブアクセスを高速化できる。

ロボットがサイトを巡回する周期

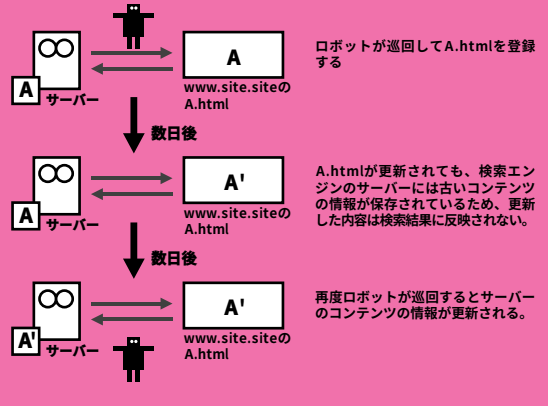
ほとんどのサイトは内容が更新されるので、過去に登録したサイトに再度ロボットを巡回させて新しいデータベースを作らなければならない。

毎日すべてのサイトを巡回することが理想的だが、サーバーの処理能力やインターネットの帯域の限界があるため実現は困難だ。そのため、ロボットを定期的に巡回させている（図❸参照）。周期の設定はそれぞれの検索エンジンで工夫しており、効率のよい更新を行っている。

検索エンジンによっては、更新頻度が高いページへの巡回周期を短くしたり、人気キーワードを含む場合は周期を短くしたり、数回ロボットを巡回させてもサイトの更新が見つからない場合は、その後の更新をストップしたりと効率化を図っている。

自分のサイトのアクセスログが見られる場合は“ robots.txt ”へのアクセス要求を確認すると、ロボットの巡回日がわかる。robots.txtとは、サイト運営者が検索ロボットの登録の可否を設定するテキストファイルだ。通例、ロボットは最初にrobots.txtを読み込もうとする。ルートディレクトリー直下のrobots.txtがある場合は、その設定に従うが、ない場合はすべてのページを登録しようとする。

図❸　サイト巡回の例



一度ロボットがサイトを登録してから、次に巡回するまでに、サイトの内容が更新されると、検索結果と異なってしまう。頻繁に更新される「ニュース系サイト」や「日記」などは検索結果と実際のサイトが異なることが多い。

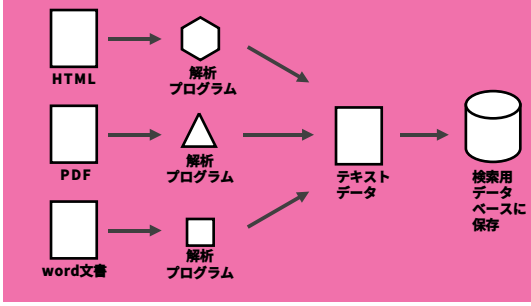
検索できるのはHTMLだけか

CGIで作っているページは、検索エンジンで登録されないことがある。掲示板やチャットのようなCGIを使って作るページの場合、アクセスごとに情報が変化することや、プログラミングのミスによる不具合の問題などを考慮して、登録しない検索エンジンもある。

ほかにも、テキストデータとして解析できる場合に検索の対象になるファイルがある。テキスト形式やPDFなどがこれにあたる。

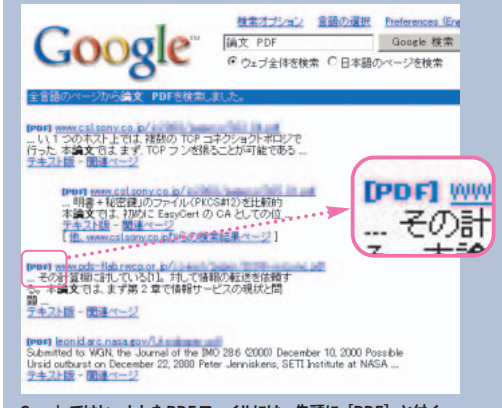
データベースにため込むだけなら、インターネット上で公開されているword文書やPDFでも技術的には可能だ。ただし、「ウェブの検索エンジン」として考えた場合、リンクのアンカーテキストがないなど、意図したテキストの解析が困難なために、これらのファイルを登録をしていない検索エンジンもある（図❹参照）。

図❹　解析プログラムと検索対象



解析プログラムを用意すれば、さまざまなフォーマットのファイルも検索できるようになる。しかし、リンクがないワープロソフトの文書のように、ウェブの検索エンジンとして必要な情報がないファイルは、あえて検索対象に含まないこともある。

Tips! Googleでヒットした PDFファイル



Googleではヒットした PDF ファイルには、先頭に［PDF］と付く

robots.txt：ロボット型検索エンジンのサイト収集プログラムに対し、登録の可否を定義するテキストファイル。「http://www.some.site/robots.txt」のように、ルートディレクトリー直下に設置する。

ロボットが集めたページはどう処理されるのか

集めたページは一度分割される

ワープロソフトや表計算ソフトにも検索機能はあるが、それらは単にファイルの先頭から順番に文字の並びを探すだけの単純な処理にすぎない。全文検索エンジンは、高速で精度の高い検索をするために、もっと複雑な処理を行っている。ここでは、その全文検索エンジンの仕組みを紹介しよう。

原田昌紀

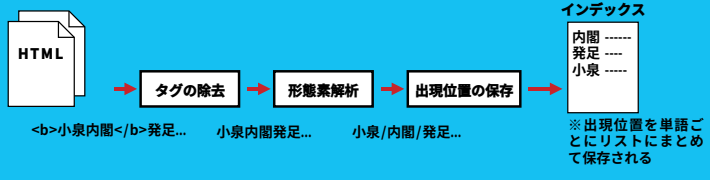
背後にある巨大なインデックス

英語の場合にかぎって言えば、全文検索エンジンの動作原理は次のようになる。

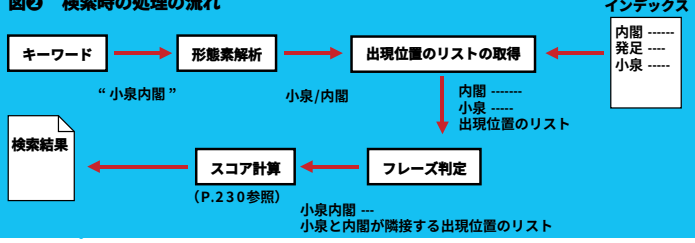
まず、ロボットで収集したHTMLテキストからタグを取り除き、残ったテキストの中の文章をスペースで区切って単語に分ける。そして、1つ1つの単語ごとに、それがどのページの何文字目に出現しているかという出現位置の情報を、リストとしてデータベースに保存する（図❶参照）。このデータベースは書籍の巻末索引を巨大にしたようなものであり、インデックスと呼ばれる。インデックスのサイズは日本の検索エンジンでも100ギガバイト、世界的な規模の検索エンジンでは1テラバイトを超えるほど巨大なものになる。

インデックスのおかげで、検索時にはキーワードの出現位置をインデックスから読み出すことで、瞬時に検索結果を得られる。また、文字そのものではなく、単語の出現する位置を求めるので、たとえば“apple”をキーワードにして検索したときに“pineapple”がヒットするといった問題が起きない利点がある（図❷参照）。

図❶　インデックス作成時の処理の流れ

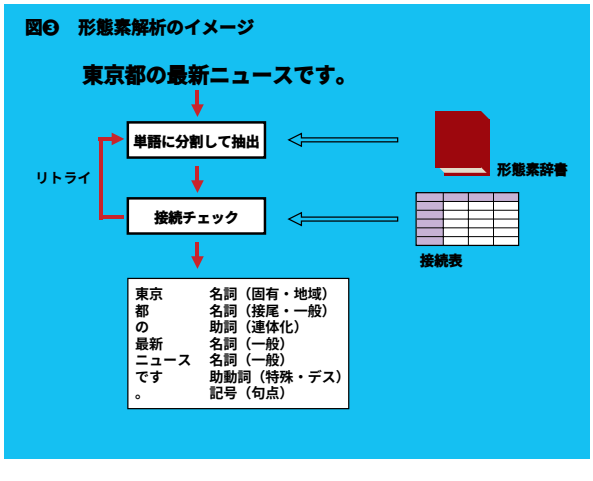


図❷　検索時の処理の流れ



形態素解析が日本語検索エンジンの要点

図❸　形態素解析のイメージ



日本語の文章は、英語などのようにスペースで区切られていないため、文章を一度、単語で区切る「形態素解析」（図❸参照）を行って、それぞれ単語を取り出してからインデックスを作成する。形態素とは意味的に分割できる最小単位の文字の並びのことで、形態素解析とは文章を形態素に分割し、その品詞を確定する処理を行うことをいう。日本語の全文検索では、入力した検索キーワードも形態素解析によって分割され、キーワードを構成する形態素と、それと同じ形態素が隣り合わせにあるかどうか（隣接判定）が検索結果として求められる。

形態素解析プログラムは、たくさんの形態素の情報を収録した辞書と、どの品詞の直後にどの品詞が出現できるか（たとえば、名詞の直後は名詞や助詞が多く、代名詞が出現することはない）という品詞間の関係を示した「接続表」を参照することで、文章をできるだけ文法的に正しい形に分割して出力する。こうした形態素解析は検索エンジン以外にも、翻訳ソフトウェアや、音声読み上げソフトウェアなどで利用されている。

全文検索：インターネットで公開されている情報の全文をデータベースに蓄積し、キーワードで指定した文字列に対応するものがあるかどうかを検索すること。ディレクトリー型検索の対照として使われる。ウェブ以外にも、CD-ROMなどの一般の文書や文献のデータベースに対する検索も全文検索方式をとる。

分割がうまくいかないと…

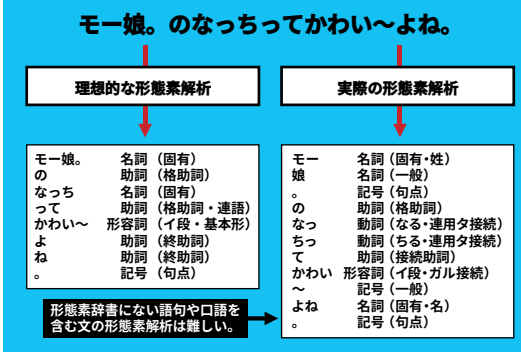
最近の形態素解析プログラムの精度は向上しており、新聞記事などが対象であれば、ほとんどの文章は文法的に正しく分割できる。しかし、サイトによく見られるような口語調の文章の形態素解析は容易ではない。

分割が正しく行われなくても、実際には問題なく検索できるケースと、まったく違う単語で検索されてしまうケースの2通りがある。右図の例でいえば、「モー娘。」[*1]というキーワードなら検索できるが、「なっち」[*2]というキーワードでは検索できない。「モー娘。」は1つの固有名詞なのに、「モー」と「娘」と「。」の3つの形態素に分かれてしまっているので形態素解析は失敗している。それでも、キーワードの「モー娘。」も同じように間違って3つの形態素に分割されるため、隣接判定でインデックスと合致できて、結果としては問題なく検索できる。一方、「なっち」というキーワードは「なっ」と「ち」に分割されるが、右例文の形態素解析では「ち」ではなく「ちっ」と認識されるため、正しく検索できない。

このように新語やひらがなばかりの単語はうまく検索されない場合があることを理解しておこう。なお、形態素解析の精度を高めるには形態素辞書に収録する形態素の数を増やすことが効果的であり、検索エンジンは次々と登場する新語に対応するように形態素辞書を常にメンテナンスしている。

*1 モー娘。：「モーニング娘。」の略。
*2 なっち：人気アイドル「モーニング娘。」のメンバー、安部なつみの愛称。

図④ 形態素解析失敗の例



記号の有無が明暗を分ける

ほとんどの記号は無視され、スペースとほぼ同様に扱われる。たとえば、「kensaku@sa-chi.ne.jp」というキーワードは、実際には「"kensaku sa chi ne jp"」というフレーズとして検索される。一方、ページの文章中の記号もスペースと見なされ、「kensaku@sa-chi.ne.jp」は5つの単語に分割されたインデックスに保持される。よって「kensaku@sa-chi.ne.jp」を含むページはすべて検索されることになる。つまり、厳密には記号は検索されていないのだが、実用上は記号を含んだキーワードでも正しく検索されているように見えるわけだ。

しかし、記号の有無には要注意だ。「クライアント・サーバー」と「クライアントサーバー」、「F-1」と「F1」では異なった検索結果が得られることが多い。

実験：記号を検索できる検索エンジンは？

通常の全文検索エンジンでは記号はスペースと同様に扱われるが、ときには記号が検索できないと困ることもある。ここでは6つの代表的な検索エンジンで、3種類のプログラミング言語、「C」「C++」「C#」を区別して検索できるかどうかを試した。検索条件は複数の単語を組み合わせた「フレーズ検索」で、記号は半角文字で入力した。フレーズを認識させるために、前後を" "で囲んだ。

結果は表に示したとおり、うまく検索できたのはGoogleのみだった。しかし、Googleでもなぜか"C#プログラミング"と"Cプログラミング"は区別できないようだ。当面はなるべく記号を使わないで済むように別のキーワードを考えて検索をするしかなさそうだ。

実験：「C」「C++」「C#」を区別できるか？

	プログラミング言語C	プログラミング言語C++	プログラミング言語C#	結果
Google	2310件	225件	14件	正しく検索できた
Lycos	2462件	87件	0件	CとC++は正しく検索できている
goo	993件	993件	993件	すべて同じ結果
Excite	1793件	1793件	1793件	すべて同じ結果
フレッシュアイ	47件	47件	47件	すべて同じ結果
infoseek	2557件	4件	2557件	C++だけほとんど検索されない

隣接判定：検索エンジンのインデックスの中に蓄積された、単語の出現位置リストを取り出したとき、検索キーワードを構成するそれぞれの単語の並び順と出現位置のリストがマッチしているかどうかを判断すること。出現位置リストは、あらかじめ単語1つ1つがページのどの位置に出現して構成しているかをデータベースにしたもの。

大文字と小文字、全角と半角の違い

多くの検索エンジンでは、アルファベットの大文字と小文字は同一視される。つまり、「windows」「Windows」「WINDOWS」のいずれで検索しても同じ検索結果が得られる。その一方で、先頭が大文字のキーワードや、すべて大文字のキーワードを特別扱いするinfoseekのような検索エンジンもある。この機能は略語や人名を検索する際に便利だ（右表参照）。infoseekでもすべて小文字のキーワードを用いると、大文字・小文字の区別なく検索されるので、必要なとき以外は小文字のみ使うようにすればいい。全角文字と半角文字はどの検索エンジンでも同一視され、区別する方法はない。

●大文字・小文字が区別される場合とされない場合（infoseekの例）

	マッチする単語		
キーワード	oda	Oda	ODA
oda	○	○	○
Oda	×	○	○
ODA	×	×	○

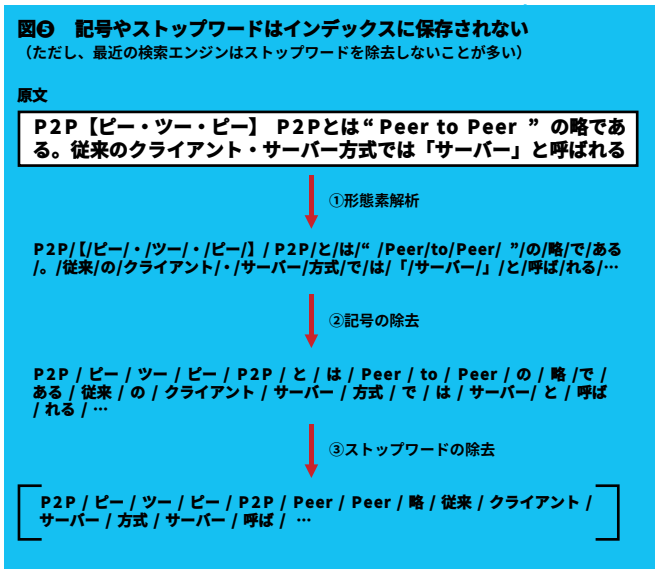
※人名（織田・小田など）と政府開発援助を意味する「ODA」でテストをした。人名で検索したい場合は頭文字だけをアルファベットの大文字にしたほうがヒットしやすい。

無視されるキーワード

全文検索ではページの中のすべての単語の出現位置をインデックスに保存するのが原則だ。しかし、助詞や助動詞は、ほとんどの文に現れるため、出現位置をインデックス上に保存するための領域を多く必要とするが、そのわりには検索に用いられることはほとんどない。そこで、こうした語句は「ストップワード」または「不要語」と呼び、出現位置をインデックスに保存しない場合がある（図❻参照）。英語の場合は、冠詞やbe動詞、前置詞などがストップワードとなる。「2」のような数字も一種のストップワードだ。

ただし、ストップワードも、フレーズの一部として使われた場合は重要な意味を持つ場合がある。たとえば、ミュージカルの「the sound of music」を検索する場合、theやofが無視されるとまったく無関係な検索結果が得られてしまう。そこで、最近の検索エンジンはストップワードも検索できるように改良されてきている。

図❻　記号やストップワードはインデックスに保存されない
（ただし、最近の検索エンジンはストップワードを除去しないことが多い）



Tips! Googleでストップワードを検索する方法

Googleでは、無視されたストップワードがある場合には、それを教えてくれる機能がある。

また、ストップワードになる単語も、検索キーワードの先頭に半角で「+」を付ければ、無視されずに検索されるようになる。ここで+と単語の間にスペースを入れてはならないことに注意しよう。

例：
+iモード
"stand +by me"

ストップワード：助詞や助動詞のように、頻繁に使われる言葉や文字は、検索の絞り込みをするときにあまり意味をなさないため、はじめから検索キーワードから省くようにしてある語句のこと。itやtheの冠詞や一桁の数字などもこれに含まれる。

「表記ゆれ」は要注意！

大文字小文字や全角半角は同一視されるが、それ以外の表記ゆれは同一視されない。たとえば、「焼きそば」を「焼そば」「焼きソバ」「ヤキソバ」と書くこともできるが、表記が異なれば検索結果も異なるのが一般的だ。

こうした問題は、検索エンジンが「同義語辞書」や「あいまい検索技術」を導入していれば解決できる。たとえば、infoseekは「スパゲッティ」と「スパゲティ」で同じ検索結果が得られる。しかし、多くの検索エンジンは表記ゆれによる検索漏れよりも、入力されたキーワードに一致するページだけを高速に検索することを重視している。検索がうまくいかなければ、別の表記や同義語の検索結果を確認する必要がある。

●表記ゆれの種類

	例
送り仮名	問い合わせ/問合わせ/問い合せ/問合せ
外来語	インターフェース/インターフェイス/インタフェイス
漢字とカナ	ねずみ/ネズミ/鼠
略語	卒論/卒業論文
漢字表記	出会い/出逢い/出遭い
同義語	老人/高齢者/年寄り/シルバー

実験：形態素解析の落とし穴

キーワードに表記ゆれがある場合、形態素解析が正しく行われたとしても、意外な形で検索漏れが起きることがある。たとえば、「フリーソフト」というキーワードで、「フリーソフトウェア」を検索することはできない。前者は「フリー」と「ソフト」に分割され、後者は「フリー」と「ソフトウェア」に分割されるためだ。

実際の検索結果から、このことを確かめられる。もし、「フリーソフト」が「フリーソフトウェア」にマッチするなら、「フリーソフトANDフリーソフトウェア」の検索結果数は、「フリーソフトウェア」の検索結果数と同じになるはずだ。なぜなら、「フリーソフトウェア」が出現するページには、必ず「フリーソフト」も出現することになるからだ。ところが、実際には大幅に少ない検索結果になる（下表参照）。

これは、形態素解析が裏目に出て「フリーソフト」と「フリーソフトウェア」が別の言葉として処理されたことを意味する。ただし、infoseekでは「フリーソフト」と「フリーソフトウェア」は同義語として特別扱いされているために問題が起きなかった。

実験：「フリーソフト」で「フリーソフトウェア」は検索できるか？

	フリーソフト	フリーソフトウェア	フリーソフトANDフリーソフトウェア	結果
Google	約127000件	約17900件	約2950件	違う言葉として認識している
Lycos	71629件	9347件	1289件	違う言葉として認識している
goo	118814件	10673件	1612件	違う言葉として認識している
Excite	170520件	15742件	2294件	違う言葉として認識している
フレッシュアイ	4112件	599件	79件	違う言葉として認識している
infoseek	79462件	79462件	79462件	同じ言葉として認識できている

Tips! AND検索のときもフレーズを意識する

たいていの検索エンジンでは、検索条件を指定せずに複数のキーワードをスペース区切りで入力すると、AND検索が行われる。しかし、単にすべてのキーワードが出現するだけのページよりも、それらが入力された順番で近接して出現するページのほうがスコアが高くなるため、フレーズ検索の場合とよく似た検索結果が得られる（詳細は230ページ）。したがって、特に検索条件を指定しない場合でも、キーワードの入力順序には配慮したほうがよい。たとえば、「システム 管理」と「管理 システム」では、検索結果が異なってくる。ちなみに、人名などを検索する場合には、半角のダブルクォート「"」を使って、厳密なフレーズ検索を指定しないと正確な検索が行われない。なお、各検索エンジンのコマンド一覧を231ページの表でまとめているので、参考にしてほしい。

あいまい検索：はっきりとしたキーワードを覚えていない状態でも、目的のものを探し出せる検索方法のこと。「単語？」「単語*」「*単語」などの入力を受け付ける検索エンジンもあれば、入力した語句から自動的に類義語などを選び出す検索エンジンもある。「ファジー検索」などと呼ばれることもある。

なぜ、このサイトが上位に来るのか

検索されたサイトを採点する仕組み

検索エンジンは全文検索によって、キーワードを含むページを何万件も検索する。しかし、いくら多くの検索結果が得られても、実際に利用者がそれらすべてにアクセスするのは非現実的だ。そこで、検索エンジンはそれぞれのページが検索キーワードと適合する度合いを「スコア」として計算し、検索結果をスコア順にランキングして表示することで、利用者が効率よく情報を入手できるように工夫している。

原田昌紀

スコア計算はハイブリッド

検索結果のランキングは検索エンジンの使い勝手に大きく影響する部分であり、検索エンジン間で激しい開発競争が行われている。そこで、最近の検索エンジンでは、さまざまな要素を加味してスコアを計算している。

まず、基本的には検索に使われたキーワードが多く出現するページほど、利用者が求めている情報を含む公算が大きいと判断され、高いスコアが与えられる（A）。

ただし、キーワードの出現回数が多くても、ページサイズが大きい場合にはスコアは低くなる（B）。多くの文字が含まれている大きなページにはキーワードが多く出現するのが当然だからだ。

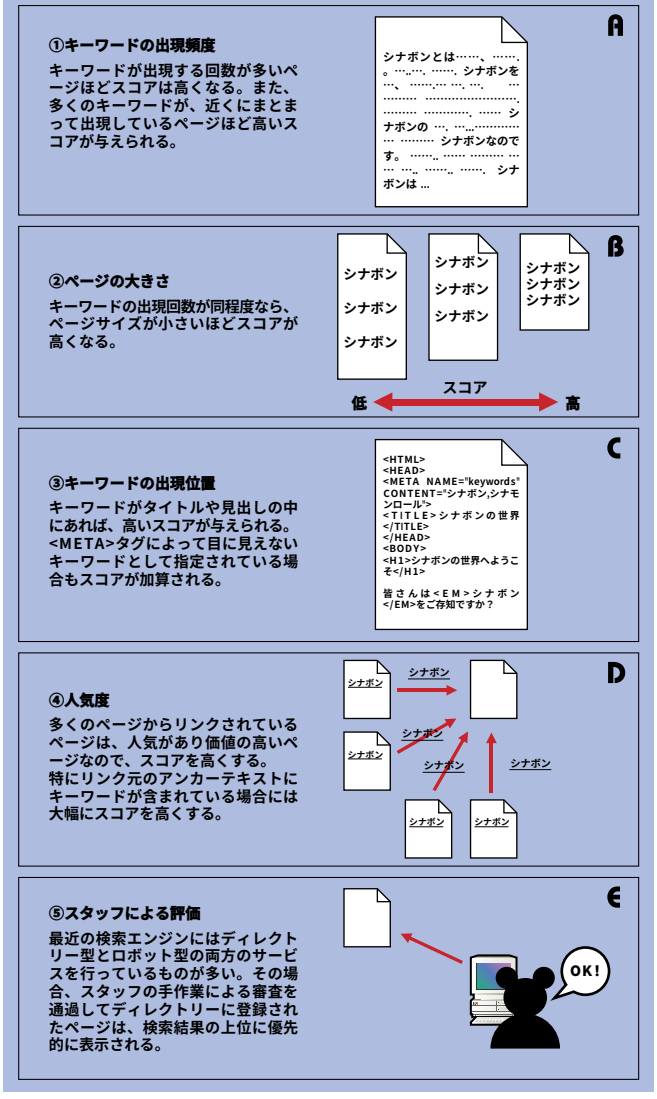
また、キーワードの出現位置もスコアの計算に使われる（C）。タイトルや見出しにキーワードが含まれているページは、本文にだけキーワードが出現するページよりもスコアが高い。本文に出現する場合も、ページの先頭付近に出現すれば少し高いスコアが与えられる。また、2つ以上のキーワードで検索した場合は、それらが入力された順番で、近接して出現するページほど高いスコアとなる。

最近注目を集めているのが、多くのページからリンクされているページを優先的に表示する方法だ（D）。これは「人気の高い（リンクを多く張られている）ページは検索エンジンの利用者にも有用だ」という考えによる。GoogleのPageRank技術もこの考えを拡張したものだ。とりわけ、リンク元のアンカーテキストにキーワードが含まれている場合、リンク先のページはキーワードと強く関連する可能性が高いとみなされ、大幅に高いスコアが与えられる。

検索エンジンのスタッフによる審査に合格してディレクトリーに登録されたサイトは、そうでないサイトよりも信頼できるものとして、優先的に表示する検索エンジンもある（E）。

最終的なスコアは、これら5つの評価項目を踏まえて総合的に決定される。このため、キーワードの出現回数だけがいくら多くても、ランキングの1位になるとは限らないのだ。

●スコアの計算方法



アンカーテキスト：<A HREF="ファイル名">..</A>というようなHTMLタグでドキュメントの特定の場所（リンク先）を指定した場合、サイトに表示させるテキストのこと。ちょうど、<A>..</A>で囲まれている部分になる。

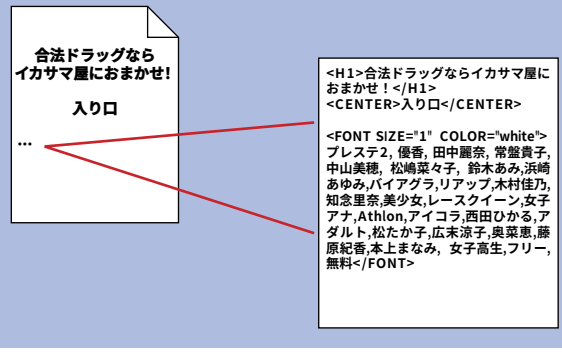
「だまし」のページがひっかかる理由

残念なことに、検索結果の上位に表示されやすいように、不正に細工されたページが少なからず存在する。このような行為は「ウェブスパム」と呼ばれている。特に多いのが、よく検索されるキーワードを背景と同じ色で大量に記述する方法だ。

このような単純なウェブスパムでも、検索エンジン側で自動判別して排除するのはそれほど容易ではない。たとえば、悪意はなくとも、背景が黒いページに白い表があり、その中に黒い文字が使われることはある。そのため、現状ではGoogleのように人気度を重視する方法がもっとも効果的な対策となっている。

なお、自分のサイトを多くの人に見てもらいたいからといって、ウェブスパムは行うべきではない。一度検索エンジンに不正行為を見破られてしまえば、二度と検索されなくなってしまう。

●ウェブスパムの例



Tips! キーワードが含まれていないページが検索されるのはなぜ？

次のような理由が考えられる。

1. ページが更新された。
検索キーワードを含んだ文章が、ページの更新によって削除された場合。ロボットがHTMLファイルを取得したあとで、そのページの内容が更新されたとしても、再びロボットが取得するまでは古い内容が検索され続ける。

2. ウェブスパム
検索エンジンをだますために背景と同じ色でキーワードを記述したページがある。

3. METAタグ
HTMLのMETAタグで検索エンジン用のキーワードが指定されている場合。このキーワードはブラウザー上では見えない。

4. アンカーテキスト
リンク元のアンカーテキストにキーワードが含まれている場合、リンク元とリンク先の両方が検索される。「inpress」というキーワードで、インプレス（正しくは「impress」）のページが検索されるのは「inpress」というアンカーテキストが存在するためだ。また、日本語キーワードで英語のページが検索されることもある。

●おもな検索エンジンのコマンドおよび機能一覧

検索エンジン		Google	Lycos	goo	Excite	フレッシュアイ	infoseek
URL		Jump www.google.com	Jump www.lycos.co.jp	Jump www.goo.ne.jp	Jump www.excite.co.jp	Jump www.fresheye.com	Jump www.infoseek.co.jp
コマンド	AND検索（両方含む）	スペース	スペース／AND	AND／&	スペース／+／AND	AND／&	+／AND
	OR検索（いずれかを含む）	なし	OR	OR	OR	｜／or	スペース
	NOT検索（含まない）	-	NOT	NOT／-	-	~	-
	AND NOT検索（右辺の語句を含まない）	（NOT検索と同じ）	（NOT検索と同じ）	（NOT検索と同じ）	AND NOT	（NOT検索と同じ）	（NOT検索と同じ）
	フレーズ検索	"" ""で囲む	プルダウンメニュー利用	"" ""で囲む	"" ""で囲む	×	"" ""で囲む
	演算式のグループ化	（）	（）	（）	（）	（）	（）
機能	大文字と小文字の区別	×	×	△ *3	×	×	△ *3
	全角と半角の区別	×	×	×	×	×	×
	自然言語認識 *1	○	○	○	○	○	○
	記号の認識	○	×	×	×	×	△ *3
	被リンク検索 *2	link : <URL>	×	プルダウンメニュー利用	×	×	link : <URL>
	備考	OR検索は検索オプションを利用	スーパーサーチで検索範囲指定や画像検索可	詳細サーチで更新日検索やデータタイプ指定可	ニュース検索や翻訳検索を用意	検索式の自動作成と保存ができる	iモードサイトやメルマガ検索、日付検索が可能

*1「おいしいラーメンの作り方」などの文章を検索キーワードとしてテストを行った。その結果、意図したページが検索されたかどうかで判断した
*2 自分のサイトなど、そのURLに向けてリンクが張られているページを検索できる
*3 編集部でテストをした結果、区別できるものとできないものがあったため

METAタグ：ブラウザーに表示される内容には直接関係なく、そのページの著作権や内容の要約、キーワードや有効期限など、そのファイルそのものの情報を記述するためのタグ。実際には、検索のためのキーワードやそのページの説明文、文字コードやリロードの指定などの記述に用いられることが多い。

「第３世代」の基本技術を理解しよう

Googleの検索テクノロジー徹底解剖

検索エンジンはインターネット普及の初期段階からいくつかの変遷を遂げており、現在は第３世代にあると言われている。ここでは「BIGLOBEサーチ」を例に、そのコアテクノロジーであるGoogle社のPageRank技術を紹介する。

高野 元／NEC BIGLOBE パーソナルサービス事業部マネージャ

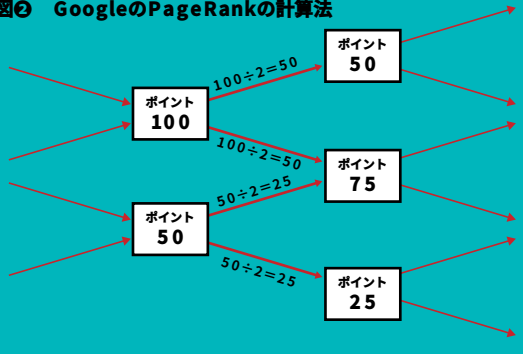
GoogleのPageRank技術

Google社は、98年にスタンフォード大学の研究プロジェクトが独立したベンチャー企業で、そのコアテクノロジーの1つがウェブ検索のランキング技術「PageRank」だ。これはウェブのリンク構造に着目した技術である（図❶参照）。

PageRankは、「他のページからどれだけリンクされているか」という参照度をもとにページの重要度を算出する（図❷参照）。このとき、人気のあるサイトからのリンク（Yahoo!など）は、他のリンク（個人ページなど）よりも重要度を上げるなどの工夫をしている。

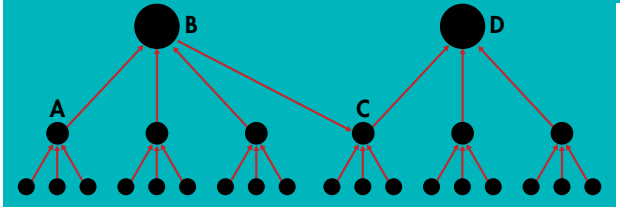
この重要度計算のアルゴリズムは、実は簡単な行列計算だが、実現には、「世界中のリンクがどのように張られているか」をロボットで収集して大規模な行列処理を行うなど、強力なシステム技術が必要となる。これもGoogle社の競争力の源泉となっている。

図❷　GoogleのPageRankの計算法



PageRankでは、ページの持つポイントをリンク数で割った数がリンクされたサイトのポイントとして計算される。このため、リンク集のような多数のリンクを張っているサイトからのポイントは低い。

図❶　ウェブサイトのモデル図



Aは3つのページからリンクを張られているので、下のページより重要と考えられる。BはAを含む多数のサイトからリンクを張られているので、Aより重要と考えられる。これを繰り返し、図の中ではDが最も重要となる。

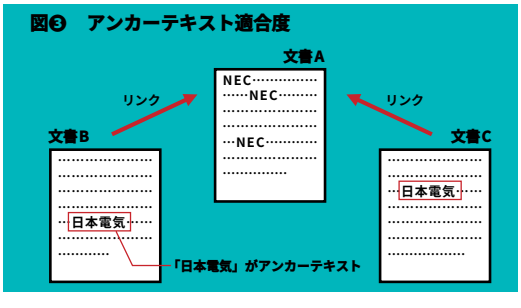
適合度を上げるために

リンク関係からページの重要度を計算しても、ユーザーのキーワードに適合しているかどうかはわからない。GoogleはPageRankに加えて、検索語と各リンク元の文字列（アンカーテキスト）との適合度を重視することでキーワードにふさわしいページを上位に出力している。

これは、アンカーテキストがリンク先のページやサイトを適切に要約していることに着目したものだ。検索結果のトップにサイトのホームページが現れるのは、この方式が大きな理由になっている。

図❸の例では、「文書A」の中には「NEC」しか入っていないが、ほかのページから「日本電気」というアンカーテキストでリンクを張られていると「日本電気」で検索しても「文書A」が高いランキングでヒットする。

図❸　アンカーテキスト適合度



文章中に「NEC」しかなくても、ほかのページから「日本電気」でリンクを張られている場合、キーワード「日本電気」で検索しても文章Aがヒットする。

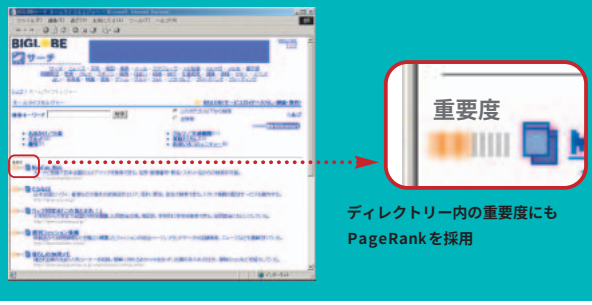
アルゴリズム：プログラム上の処理の順番や手順、方法のこと。

ディレクトリーへの適用

「ディレクトリー」も検索に役立つが、1つのカテゴリー内に数十ものサイトが登録されていると判断基準がほしくなる。BIGLOBEサーチでは、サイト表示順序にもPageRankを利用することで重要なサイトを上位に表示している（図❹参照）。

BIGLOBEサーチでは、PageRankを採用したことによって、検索結果に対するユーザーの満足度が向上している。PageRankによる検索結果の向上は、iモードなどの携帯電話でのインターネットでも、さらに大きな力を発揮するものと思われ、今後の大きな流れになるだろう。

図❹　ディレクトリー検索画面の例



ディレクトリー内の重要度にもPageRankを採用

第3世代検索エンジンの特徴は

第1世代の検索サービスを「人手によるサイト収集・分類」、第2世代を「ロボットによる自動ページ収集」とすると、第3世代検索サービスの特徴は「検索機能の統合」と「ページのランキング技術」にあると言える。

「BIGLOBEサーチ」は「キーワード1つで探している情報へ」をコンセプトにして、「ページ検索」「カテゴリー検索」「的中ナビ検索」を統合し、またGoogle社のPageRank技術を利用することで、探しているサイトを簡単に見つけ出せるようにしている。

図❺　キーワード検索画面の例

ページ検索ではPageRankとアンカーテキストの適合度でポイントを計算

Jump search.biglobe.ne.jp

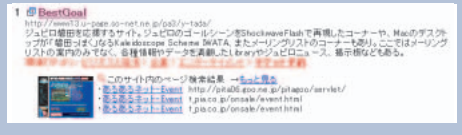
「goo」のロボット型ディレクトリー「ハイブリッド型検索エンジン」とは

一般的に「ディレクトリー型」と「ロボット型」に分けて語られてきた検索エンジンだが、くわしく分類すると、①サイトの収集方法が「人手」か「ロボット」か、②検索方法が「ディレクトリーたどり型」か「キーワード入力型」か、③出てくる結果が「サイト」単位か「ページ」単位か、という3つのレベルでの境界があった。

「goo」を運営するNTT-Xは「ハイブリッド型検索エンジン」を開発した。人とロボットの協働で、「ディレクトリーをたどる」検索方法でも、キーワードを入力する検索方法でも、「カテゴリー検索」と「サイト検索」と「ページ検索」の結果を同時に得られるようになる。この「ハイブリッド型検索エンジン」サービスは、7月14日から「goo」で提供する予定。　（NTT-X　鈴木基久）

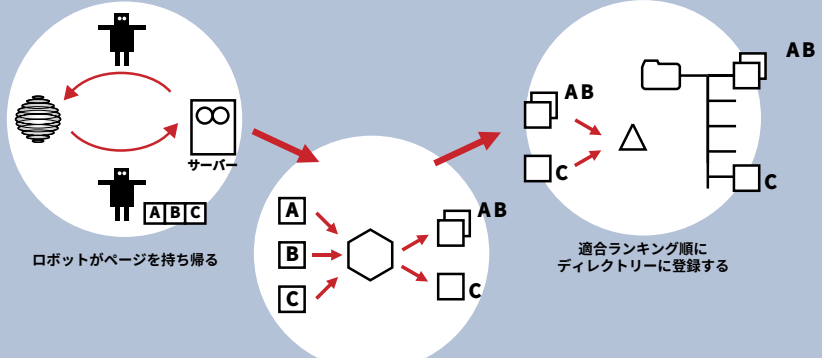
Jump www.goo.ne.jp

●縮小版入りの検索結果



ヒットしたサイトのほか、関連するカテゴリーや、サイト内の各ページ、ヒットしたページの縮小版が表示される。

●ロボット型ディレクトリー



ハイブリッド：2つの異なる要素を融合させること。

統合検索、代理検索がキーワード

これから注目の次世代の検索テクノロジー

ここまでは、現行の検索エンジンの主要テクノロジーを中心に解説してきたが、これからの検索エンジンの方向性を感じさせるサービスを２つ紹介しよう。
編集部

画像や音楽も検索「ネイバー」

ネイバーは、韓国のNAVER.COM社が運営する「NAVER.COM」の日本語版で、自然語検索、Q＆A検索、マルチメディア検索が特徴の検索エンジンだ。

ある単語を入力すると、予想されるサイトをQ＆A方式で誘導して、検索エンジンに慣れていない人でも目的のサイトに到達できるようになっている。

たとえば有名人を検索した場合、プロフィールが表示され、公式サイトのほか、有名なファンサイトがプルダウンメニューで表示される。

また、Altテキストなどを参考に、キーワードに合う画像やサウンドがネイバーのデータベースに登録されているときは併せて表示される。

Jump www.naver.co.jp

❶Q&A検索

❷画像検索

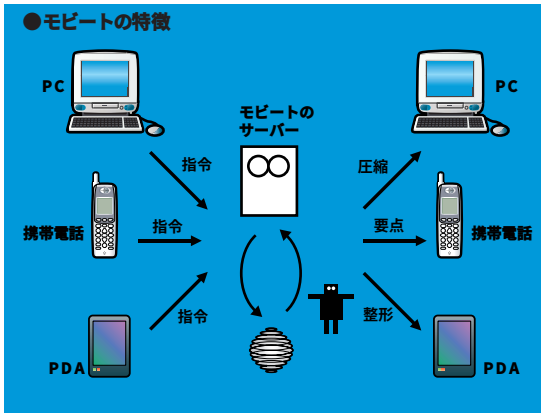
❸サウンド検索

❶Q＆A検索では、ユーザーの探そうとしているサイトを先読みして、質問を出して目的のサイトへ誘導する。

❷キーワードに合う画像も検索できる。

❸リアルオーディオなどのほか、「MP3」ファイルの検索も可能だ。

●モビートの特徴



モビートでは、指令を出す端末と情報を受ける端末が別でもOKだ。たとえば、携帯電話から検索するキーワードを指定して、携帯電話をインターネットから切断、あとから検索結果をPCで受け取ることができる。

検索代理人「モビート」

モビートは、ユーザーの代わりにページを検索する「検索エージェント」（代理人）だ。エージェントはモビートのサーバーに置かれるため、インターネットに接続していないときでも、入力した語句にヒットするページをモビートが検索し続けてくれる。

ユーザーがモビートのサーバーに指令を送ると、モビートはサーバー上でウェブの検索を開始して、検索結果をサーバーに蓄積する。ユーザーが再びサーバーにアクセスし、モビートにコンテンツを持ち帰るように命令を出すと、モビートはユーザーが使う端末に合った形で検索結果を持ち帰る。

たとえば、モビートをPCに呼び戻す場合は、検索結果としてヒットしたサイトのHTMLや画像ファイルなどを圧縮して持ってくるので、効率よく受信できる。携帯電話で受信する場合は、携帯電話の表示能力に合わせた、要点だけをピックアップする。

Jump mobeet.exrd.nacsis.ac.jp

Altテキスト：HTMLにおいて、画像などが表示できないときに代わりに表示されるテキスト。

▼次世代検索エージェント「モビート」を使ってみよう

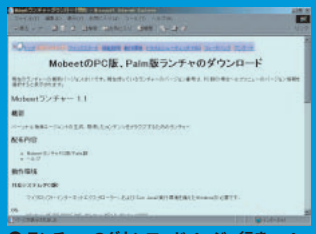
モビートのインストール

モビートの「ランチャー」（操作プログラム）は、インターネットエクスプローラ4以上がインストールされたウィンドウズ95/98/Me/NT/2000のほか、iモード、Palmに対応している。ここでは、ウィンドウズ版のランチャーのインストールと使い方を説明しよう。

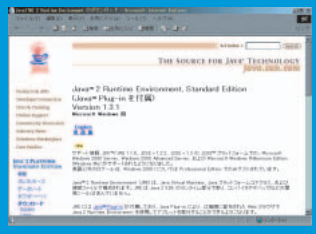
現在、モビートは携帯電話とPC、PDAの3つに対応している。街中で気になったニュースがあったら、携帯電話からモビートに指令を出し、帰社後にPCでじっくりと読むといった使い方ができる。

モビートへの指令は自然言語で行える。一度指令を出すとモビートはサーバーに送られるので、回線を切断してもかまわない。10分ほどすると、モビートが見つけたコンテンツを持ち帰る。

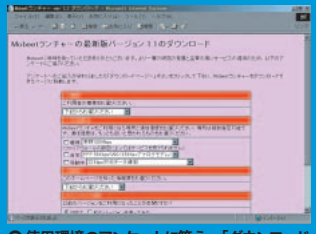
なお、ランチャーとサーバーはTCP/IPポートの9001番を使って通信するので、ファイヤーウォールのある企業や、ルーターを介した接続ではモビートを使えないことがある。



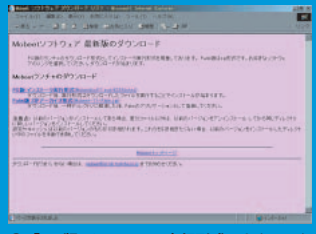
❶ランチャーのダウンロードページへ行き、ページ下部にある「Java2 Runtime Environment Standard Edition」をクリックする。



❷「Java 2 Runtime Environment, v 1.3.1」の「国際化版」をダウンロードする。

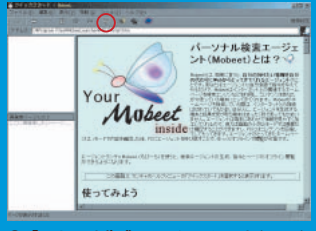


❸使用環境のアンケートに答え、「ダウンロードページへ」というボタンをクリック。

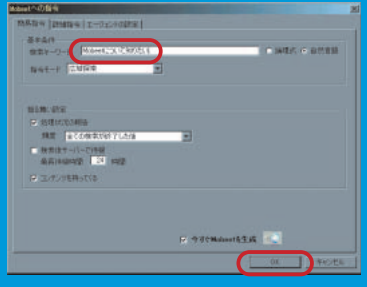


❹「PC版 インストーラ実行形式」をクリックして「MobeeInst11.exe」をダウンロードする。その後、ファイルをインストールする。

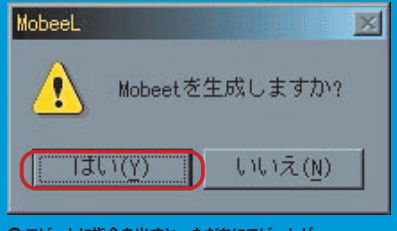
モビートランチャーの使い方



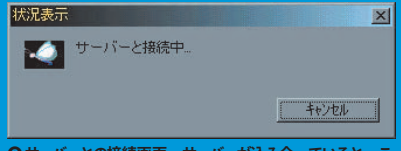
❺「Mobeetを生成」というアイコンをクリックする。



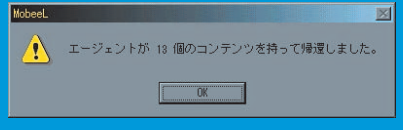
❻モビートで調べたい語句を入力し、「OK」をクリックする。



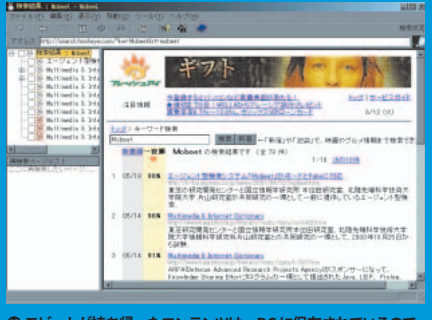
❼モビートに指令を出すと、ただちにモビートが生成される。インターネットに接続されているのを確認したら「はい」をクリックする。



❽サーバーとの接続画面。サーバーが込み合っていると、この画面が長く続くことがある。



❾モビートが持ち帰ったコンテンツの報告画面。「OK」をクリックするとダウンロードが自動で始まる。



❿モビートが持ち帰ったコンテンツは、PCに保存されているので、インターネットに接続していなくても見られる。

モビートの注意点

モビートはTCP/IPポートの9001番を使って通信するので、ファイヤーウォール内やNATを使うルーターを介したインターネット接続では利用できない。この場合は、ルーターを静的NATにして対応する。

TCP/IPポート： TCP/IPプロトコルで設置されている仮想的な端子の番号。TCP/IPポートの80番はHTTP、21番はFTPというように定義されている。サイトをブラウザーで見る場合、80番のポートを使いサーバーと通信している。

［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