

第9回 検索上手になる

何か調べたいことがある時には、GoogleやYahooなどで検索してみるのが常になりました。

世界のWebサイト数は2007年秋に1億を超えたと
言われ、インターネットは情報の宝庫から情報の海原
と呼ばれるようになりました。その中から目的の情報を
探し出すための強力な道具が、検索エンジンです。
今回は、インターネットの検索エンジンについて解説
します。



【今回登場するキーワード】

- 「検索エンジン」
- 「AND 検索」
- 「OR 検索」
- 「NOT 検索」
- 「フレーズ検索」
- 「SEO」
- 「SEM」

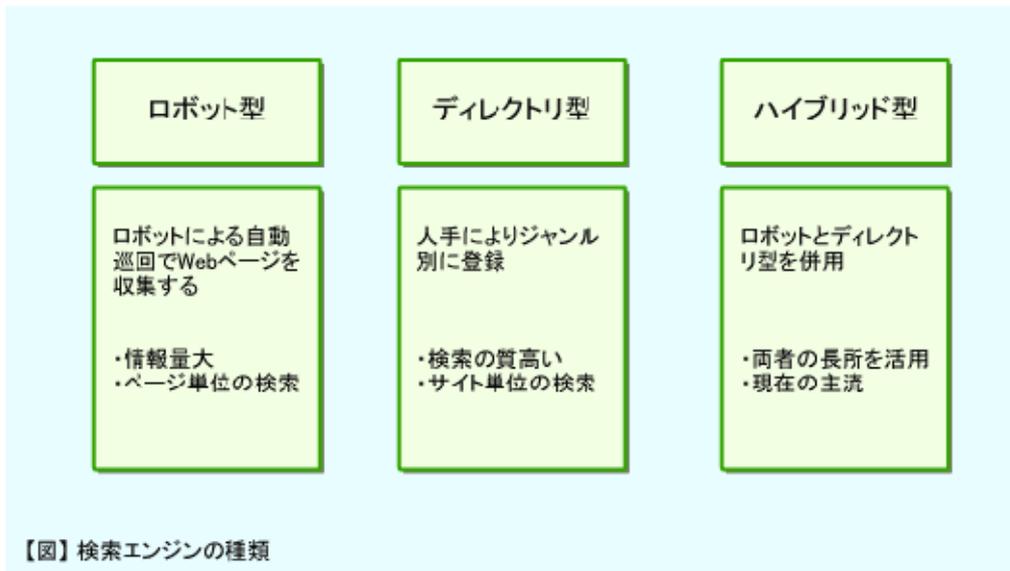
検索エンジンは、1994年4月に米スタンフォード大学の学生だったデビッド・ファイロとジェリー・ヤンが作った「Yahoo! Internet navigational guide」(後のYahoo!)が最初です。その後1995年末に「AltaVista」をDEC(現在はヒューレット・パカードに統合)が公開。1998年9月には、ラリー・ページとサーゲイ・ブリンの2人によってGoogleが共同設立されました。MSN(Microsoft Network)サーチのベータ版は2004年に登場しています。

■検索エンジンの種類としくみ

GoogleやYahoo、MSNサーチなど検索を専門にするWebサイトを検索サイトと言います。インターネットで検索上手になるためには、まず検索サイトの検索の仕組みを理解することから始めましょう。ここで実際に検索を行うプログラムが検索エンジンです。この検索エンジンは、私たちが入力したキーワードを基に数十億ページの中から該当するページ候補を見つけてくれるのです。

1. 検索エンジンの種類

検索エンジンは「ディレクトリ型」、「ロボット型」「ハイブリッド型」に大別することができます。



ロボット型

クローラと呼ばれるロボット（実体はプログラム）が Web サイトを巡回し、Web ページの情報を収集し、あらかじめ検索用の索引ファイルを作成する仕組みです。索引化するための情報収集が自動化されているために大量の情報を集めることができますので、情報量が圧倒的に大きい利点があります。Google は 80 億ページ、Yahoo! は 190 億ページを索引化していると言われています。情報が多いために、検索にあたってはキーワードを工夫しないと、目的の情報にたどりつけない場合があります。

今日では大規模な検索エンジンはすべてロボットによる Web ページ情報の収集を行っています。ロボットとして有名なものに Google の「Googlebot」、Yahoo! の「Yahoo! Slurp」、MSN の「msnbot」があります。

ディレクトリ型

Web サイトの所有者などからの要請に基づいて、検索サイトの運営者の手によって、ジャンル別に検索サイトに登録され、ユーザーは登録された中から検索する仕組みです。Web ページの登録にあたって、ある程度選別されるため、検索されることだけを目的とする Web ページやスパムなどいわゆる検索ノイズが取り除かれて目的のページを見つけやすい、企業名や地域名などで検索することで関連情報も同時に入手できるなどの利点があります。ロボット検索と比較すると、情報量では劣りますが、検索結果の質の点で優れています。検索するページのタイトルやジャンルが明確な場合には、キーワードと関連性が高い Web ページが見つかりやすい特長があります。ディレクトリ型の代表が Yahoo カテゴリです。

ハイブリッド型

ディレクトリ型とロボット型を併用した検索エンジンのことです。

ディレクトリ型検索で見つからなかった場合に、そのキーワードを引き継いでロボット検索を行う「引き継ぎ検索」のように、2つの方式を合わせた検索エンジンです。

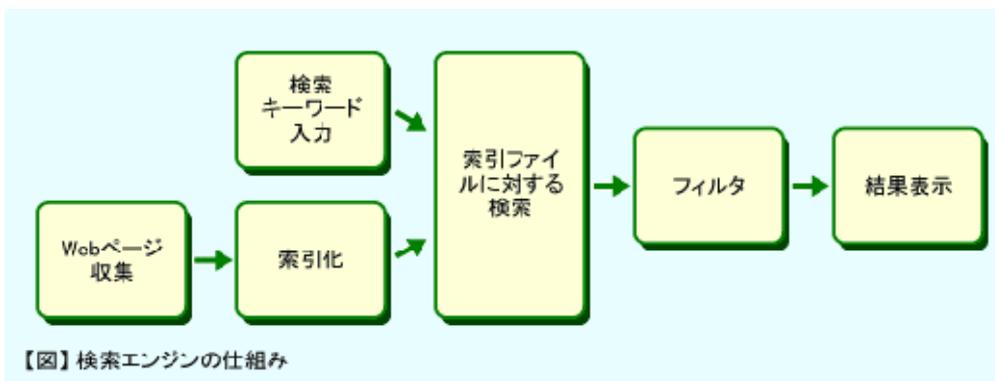
ディレクトリ型が備えている精度の高い検索機能とロボット型の情報量を組み合わせることで、より目的に近いページにたどりつき易くなります。

ハイブリッド型の代表が、NTT レゾナント（株）の検索サイト「goo」です。

ディレクトリ型検索エンジンの代表である Yahoo! JAPAN は goo と提携し、ディレクトリ型検索で検索結果がゼロになったときには、自動的に goo で検索した結果が表示される仕組みになっています。

2. 検索エンジンの仕組み

今日の検索エンジンは膨大な Web ページを一つひとつ開いて直接検索するわけではありません。あらかじめ Web サイトを巡回し、Web ページの情報を収集し、次にその情報から索引ファイルを作ります。検索はこの索引ファイルに対して実行します。さらに、実行結果を並べ替えたり削除したりする操作を行って検索結果として表示します。



上記は基本的な検索エンジンの仕組みです。この仕組みで検索した結果をそのまま表示してもユーザーの望む結果になるとは限りません。例えば、検索されることだけを目的とした Web サイトやキーワードを羅列した意味のないサイトや、時には有害なサイトもあります。ユーザーが本当に望む検索結果を表示することが、各検索エンジンの技術とノウハウの蓄積です。その結果、Web ページの収集から結果の表示に至るまで、それぞれが複雑なアルゴリズムや膨大なプログラムで処理されています。

現在最も利用されているロボット型検索エンジンを例に、個々の基本的な動作を説明します。

Web ページの収集

ロボットが Web サイトを巡回し、Web ページ内の情報から、検索される語句など索引情報を作るのに必要な情報を収集します。この作業はクローラ（Crawler）とかスパイダー（Spider）とよばれるロボット（実際はプログラム）が行っています。ロボットは Web ページ内の HTML ファイルを読んで次々にリンクをたどっていくのです。

Web ページは膨大な数があります。しかも更新の頻度や情報の質もまちまちですから、すべてのサイトを同じ間隔で巡回するのは効率的ではありません。更新頻度の高いサイトは短い期間で巡回し、更新頻度が低いページは間隔を長くしたり、検索される可能性が高い Web サイトや重要

なサイトからのリンクが多い Web サイトは巡回の優先度を高くしたりといった調整が行われています。また、検索エンジンによっては、スパムフィルターにより有害なサイトや無意味なサイトを検索対象から除いているものもあると言われます。

索引化

Web サイトで収集されたページの情報を基に、索引ファイルを作ります。この索引ファイルに記録された文字列に対して検索を行います。

索引ファイルには、索引文字列、ファイルの場所、更新日、出現頻度などが表形式で記録されます。この索引ファイルの出来によって検索の速度や精度が変わってきます。検索用の文字列の作り方で主流となっているのは、「形態素解析」と「N-gram 法」の2つの手法です。

形態素解析は、文章から検索用の語句を見つけ出すために、文脈を解析し、単語に分解して、単語単位で索引を作成します。そのためには解析用の辞書が必要です。特殊な用語、新しい言葉、造語など解析用の辞書に登録されていない単語については検索できない可能性があります。形態素解析は日本語なら日本語専用というように特定の言語を対象とするため、複数の言語で記述された文書も苦手です。

一方の「N-gram 法」は、検索対象を単語単位ではなく一定の文字数（N 文字）単位で機械的に分解します。単純に文字列片に分解するために検索の漏れはありません。解析用辞書も不要で、複数の言語で記述された文章も問題ありません。しかし、形態素解析による索引と比べると、意図したものとは異なる検索結果（検索ノイズと言います）を生じることが多く、索引ファイルのサイズも大きくなる欠点があります。

それぞれの手法に得手不得手があります。今日ではこの両者を組み合わせた方法で索引化されているようです。

検索実行

作成された索引ファイルに対して、入力されたキーワードと索引化された文字列を比較します。比較する方法には、表の先頭から順に比較する線型探索や、ソートされた表を半分ずつ比較する2分探索などがあります。前者はどんな表でも探索できる反面で検索に要する時間が長くなります。後者は高速ですが、索引がソートされていないと探索できません。

フィルタ

検索した結果をそのまま表示すると、キーワードを羅列しただけの無意味な Web サイトが上位に表示される可能性があります。そこで、検索エンジンは独自の順位付けで作為的に表示を変更します。検索結果から、アダルト関連や検索されることだけを目的にした Web サイト、スパムなどは削除されたうえで、表示順を決めるためにランク付けをします。更新日付順や文書中のキーワードの出現頻度の順、HTML の見出しタグ（<title>タグや<h1>タグなど）の比較、出現頻度や出現する文書数による単語重要度評価によって検索結果の表示順位が決まります。

Google はこの表示順位に独自のページランクという重み付けを行っています。「良質の Web サイトからリンクされる Web サイトは良質である」という考え方です。「数多くリンクされている」だけでなく、「数多くのサイトからリンクされているサイト（＝良質なサイト）からリンクされて

いる」から良質のサイトであるとし、ランクが上がります。その結果表示順位も上がります。ページランクは、ブラウザに Google ツールバーをインストールすると表示されます。数値で示されるので、判りやすく、客観的に見えるため、Web サイトの運営者は気にせずにはられません。今日の Google の人気は検索エンジンの優秀さもありますが、こうしたユーザーに判りやすい技術や手法にも支えられています。

■ 検索エンジンの使い方

検索エンジンの使い方は難しくはありません。しかし、工作中的の調べものや分からないことを調べるときに、なかなか目的の結果を得られなくて困った経験は誰しもあるはずです。そうしたときに、ほんの少し検索エンジンの機能を知っているだけで、解決できることがあります。ここで、検索エンジンの持つ基本的な機能とその使い方を説明しておきましょう。

1. 検索エンジンの機能を知る

まず、基本的な検索方法を説明します。

現在キーワード一語で目的の Web ページが検索できることはほとんどありませんので、複数のキーワードの組み合わせで使うことが多くなっています。以下の表は Google の例ですが、複数のキーワードの入力の仕方でも検索方法を切り換えることができることを示しています。スペースで区切れば AND 検索となり、スペースの後に記号を付けることで次のような検索ができるようになります。

種類	説明	入力例
AND 検索	AとBの両方のキーワードを含む	インターネット 講座
OR 検索	AかBどちらかのキーワードを含む	インターネット OR 講座
NOT 検索	Aを含むもののうち、Bを含まないもの	インターネット - 講座
フレーズ検索	「インターネット講座」のように複数の語からなる語句を一語として検索	"インターネット講座"
曖昧検索しない	検索エンジンは表記のゆれを自動的に吸収し、例えば「バイオリン」で検索すると「ヴァイオリン」も検索されます。 「バイオリン」だけを対象にしたい場合の検索方法	楽器 + バイオリン

表：検索の種類

2. キーワードを考える

検索は、「これがベスト」とされる確立されたキーワードの選択手法はありません。しかし、検索の方法を工夫することで効率を上げることはできます。例えば、複数のキーワードで検索するときには、AND検索だけでなく、NOT検索やフレーズ検索を利用することです。NOTは除外するキーワードとして例えば「-トラックバック」とすると表示結果からブログのほとんどを取り除くことができます。また「-pdf」とすれば、検索結果からPDFを除外することができます。フレーズ検索では、「インターネット講座」のように、膨大な検索結果が予想される「インターネット」というキーワードが含まれている場合に、全体を一語で検索するだけで、検索結果は大幅に絞り込まれます。

また、キーワードの選択にあたって、オンライン版の専門用語事典や類語辞書を使ってキーワードを調べ、その中に登場する別の語句と一緒に検索するのも有効です。

また、検索エンジンによってキーワードの扱いは異なります。別の検索エンジンを試してみることが有効な場合もあります。キーワード選択のコツを習得するには、豊かな想像力と発想、そして根気良く経験を積むこと以外にありません。

3. 特殊な検索

Googleを例に、いくつかの特殊な検索方法を紹介します。

ある特定のサイト内のページを見たい場合、例えば、「富士通」のプレスリリースを検索したい場合には、キーワードに「プレスリリース site:jp.fujitsu.com/」と入力します。「site:」の後に続けてドメイン名を入力することで、そのサイト内だけを対象にした検索を行うことができます。

サイト指定はキーワードの最後に入力します。この検索方法をサイト内検索と言います。

パワーポイント資料だけを見たい場合には、「SOA filetype:ppt」のようにファイルの種類を指定して「filetype:」の後にファイルの拡張子を指定します。ファイルタイプ指定検索と言います。指定可能なファイルの種類は以下の通りです。

Adobe Acrobat PDF (.pdf)
Adobe Postscript (.ps)
Microsoft Word (.doc)
Microsoft Excel (.xls)
Microsoft PowerPoint (.ppt)
Rich Text Format (.rtf)

表：ファイルの種類

さらに、Googleでは特殊な検索を指定するためのキーワードが用意されています。次の表に示します。

キーワード	説明	使用例
intitle:	Web ページのタイトルを検索します。HTML の<title>タグで囲まれた文字列です。ブラウザのウィンドウ最上部に表示される文字列	intitle:ソリューション
allintitle:	すべてのキーワードがタイトルに含まれる Web ページを検索	allintitle:SOA SaaS
inurl:	Web ページの URL 文字列を検索	inurl:fujitsu.co.jp
allinurl:	指定されたすべてのキーワードが URL 文字列に含まれるページを検索	allinurl:fujitsu usa
inanchor:	Web ページ内のアンカーテキストだけを検索。アンカーテキストとは、リンクが設定された文字列のこと	inanchor:サービス
allinanchor:	すべてのキーワードがアンカーテキストに含まれるページを検索	allinanchor:サービス サポート
intext:	WEB ページの本文の文字列（タイトル、URL、リンク部分以外）だけを検索	intext:サービス
allintext:	すべてのキーワードが本文のテキストに含まれるページを検索	allintext:サービス サポート

表：Google で特殊な検索を指定するキーワード

4. 消えた Web ページの検索

消えていった Web ページを検索することもできます。これには2つの方法あります。1つ目は、Google の検索結果に表示されるキャッシュです。既に削除されたり閉鎖されたりした Web ページも Google のロボットが巡回し収集した結果が残っているのです。それがキャッシュです。Web ページを開くとキーワードがハイライトされた状態で表示されます。2つ目は、過去に公開されていた Web ページを記録として保存している Web サイト、「インターネットアーカイブ」です。米国の非営利組織インターネットアーカイブが運営するサイトが有名です（URL <http://www.archive.org/>）。1996 年以降の数百億ページを保存しており、米国議会図書館やスミソニアン博物館も協力したと言われていています。いつの間にか消えてしまった Web ページやサイトが保存されている可能性があり、過去に遡って Web ページを探したい場合に便利です。ただし、現状では、無断で Web ページを保存しているとして著作権上の問題や、何らかの理由で非公開とされた Web ページが公開されてしまうといった問題もあり、存在に異を唱える人も少なくありません。

5. ジャンル別に検索

特定のジャンルに強い検索エンジンや最新のニュースから検索できるなどの特徴を持つ検索サイトがあります。Google や Yahoo! もオプションとして同様の機能を持っています。その中からいくつか紹介しておきましょう。

ジャンル	検索エンジン・サイト名	特徴
動画	Fooooo http://www.fooooo.com/	1億3000万の動画から検索。You Tube やニコニコ動画など動画共有サイトの動画を検索できる
ニュース	フレッシュアイニュース http://www.fresheye.com/	検索ポータルサイトのニュースのページ。最短5分前に更新されたニュース情報も探し出すことができる
ブログ	アクセラナビ http://www.accelanavi.com/	国内最大の収集記事数・サイト数を誇る国産技術によるブログ検索エンジン。RSSの更新情報だけでなく、ブログ記事全体を対象に検索が可能
写真	Yahoo!検索 (画像検索) http://search.yahoo.co.jp/images?ei=UTF-8&p=	Yahoo!の画像検索。写真共有サイト flicker の3億枚に上る写真から、投稿者や閲覧者が付けたタグで検索できる
書籍	http://www.books.or.jp/	書籍流通大手トッパンが運営する書籍検索専用のポータルサイト。出版社から提供された書籍情報から入手可能な既刊分、約80万点の書籍データベースから検索できる。
書籍	Google ブック検索 http://books.google.co.jp/	Googleの書籍検索。キーワードで書籍を検索できる。出版社の許諾を受けた書籍については、表紙や本文の一部の画像も閲覧できる。著作権が切れた書籍は全文閲覧できる

表：ジャンル別検索エンジン

■ 検索されるには

今度は検索される側から検索について考えてみましょう。

1. 検索されやすいサイト

調べものをするために検索してみたら、表示結果はブログばかりと言う経験はないでしょうか。ブログは構造的に検索されやすい特徴があります。ブログではほとんどの場合、HTMLタグを自分で入力することがなく、ブログのシステムまかせになります。これが、W3C*1の仕様に沿った標準的なHTML構造となることが多いために、検索エンジンのロボットには分かりやすいのです。また、トラックバックなど手軽に相互リンクが可能で、このことが他からリンクされていることを重視するロボットに評価されます。Googleのページランク*2は、これを数値化したものです。通常のWebサイトに当てはめると、W3C仕様に準拠した標準的なHTML構造にすること、他のサイトからリンクを張ってもらうこと、コンテンツを充実させ、ユニークなキーワードを設定しておくことが、ブログ並に検索されやすいサイトへの王道と言えます。

*1 W3C: The World wide web Consortium. Webの標準化団体。HTML、XMLなどの規格をとりまとめている。

*2 検索エンジンの Google が採用している、Web ページの重要性を測るアルゴリズム。ウェブサイトの人気度を 1 から 10 の数字で表す。Google Toolbar をダウンロードしてブラウザにインストールすると表示させることができる

2. SEO と SEM

Web サイトに 1 人でも多くの訪問者を迎えるためには、検索エンジンによる検索結果の表示順はとても重要です。そこで、検索エンジンのロボットに分かりやすいサイト作りを目指すことになります。その検索されやすいサイト作りの技法が、**SEO(Search Engine Optimization : 検索エンジン最適化)**です。検索エンジンのロボットによる巡回では、Web ページ内部でキーワードがどう扱われているかを判断して索引化します。例えば、見出しとして扱われているか、強調表示されているか、出現頻度の高低はどうか、それによってそのキーワードがページにとって重要かなどを判断します。ですから、こうした検索エンジンの情報収集の特徴に合わせてコンテンツを作れば表示順が上がります。また、Google が独自に採用するページランクを上げるために他のサイトからリンクを張るといったことも行います。

さらには、検索エンジンをもっと積極的に活用する例もあります。**SEM (Search Engine Marketing)** は、検索エンジンを広告の手段として使用するマーケティングの手法です。特定のキーワードが入力された時に、検索結果の画面に広告として表示されるキーワード連動型広告や、スポンサーリンクとして検索結果の上位に表示されるようにする有料のリスティングサービスが利用されています。

特定のキーワードに関心を持ってクリックしてサイトに来場する訪問者は、そのキーワードに関するサービスや商品に関心を持っており、ビジネスの成果に結びつきやすいと言われています。個人が手軽に情報を発信できるブログにも広告を掲載するところが多く、SEO は大きな関心事です。

3. 検索エンジンの問題点

検索エンジンの影響力は大きく、検索結果で上位に表示されることで例えば、ネット通販サイトの業績向上に結びつく可能性も小さくありません。そのため、SEO や SEM ビジネスは隆盛を極めています。SEO、SEM をキーワードに検索すると 200 万件以上という膨大な結果になります。個人のサイドビジネスからコンサルタント、企業にいたるまで SEO や SEM のノウハウを扱っています。

しかし、こうした検索エンジン頼みのマーケティングは危険もはらんでいます。ユーザーの側から見ると、必ずしも商品を購入するために検索するわけではありません。従って検索結果が通販サイトばかりでは、やがてその検索サイトが使われなくなってしまう。さらに、急増するフィッシングなどの詐欺サイトや有害なサイトは、可能な限り表示されないようにしなくてはなりません。

そこで、検索エンジンは、行きすぎた SEO や SEM で作られた Web ページや有害な Web ページを検索対象から除外するため、常に検索エンジンの改良・更新を続けています。この改訂作業の結果、アルゴリズムや判断基準のわずかな変更で、突然、検索対象外になることがあります。例えば、見出しの文字や強調文字が背景と同じ色になっているとキーワード隠しと見なして検索対

象外とする検索エンジンがあります。アクセスを増やすために、Web ページの内容とは無関係な人気のあるキーワードを画面では見えないようにしている悪質なページと判断するのです。デザイン処理で背景と似た色を使っていたり、色指定を誤っていたり、文法ミスがあったりすることで、意図的ではなかったのに、突然検索対象外になってしまうこともあるのです。

SEO・SEM 対応サイト、有害サイト、検索エンジン、この三者は常に知恵比べをしています。検索エンジンのユーザーにとっては探しやすい安全に便利になりますが、ネット通販サイトにとっては死活問題です。悪質な業者も同様です。検索エンジンをめぐって激しいせめぎ合いが続いています。

4. 検索されないようにする

逆に検索してほしくない場合もあります。例えば、グループ仲間などで運営するクローズドな Web サイトや特定の人向けにサービスを行うサイト、または作成途中のサイトなどでは、検索サイトにより無関係な人が来ることは望んでいません。こうした場合には、Web サーバのトップページと同じディレクトリに、robot.txt というテキストファイルを配置するのが一般的です。検索エンジンのロボットは、robot.txt を発見すると、ページ情報の収集を行いません。内容は次の例のような簡単なものです。

User-agent: *	ロボットを指定、「*」は全ての意味
Sitemap: http://www.xxxxxxx.xxx/sitemap.xml	検索エンジン用の XML ファイルです。
Disallow: /home/	索引化してほしくないページ

図 : robot.txt の例

robot.txt はすべてのロボットに対して有効と言うことではありません。中には紳士的でないロボットがいるかもしれません。

また、メタタグを使う方法もあります。メタタグとは HTML の中であってファイルの属性を記載するタグです。このタグに次のように記載することでロボットが情報収集をしなくなります。

```
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">
```

これも robot.txt と同じで完全に検索対象外になると言うことはありません。なお、robot.txt、メタタグの利用ともに、セキュリティ対策になるものではありません。セキュリティ対策は別の方法で行っておく必要があります。

■ これからの検索サイト

キーワードによる検索ではなく、日常使用する文章で検索する方法の開発が進められています。自然言語検索という方法で、例えば、ファミリー会に入会するための手続きを知りたい時に、「ファ

ミリ会」、「入会」、「手続き」とキーワードを連ねるのではなく、「どうすればファミリー会に入会できますか」といった質問を入力すると、答えになる Web ページが表示される仕組みです。Google キラーと呼ばれる米国 Powerset や英国 TrueKnowledge が有名で、まだ開発段階と言ったところで、これからの技術として注目されています。

しかし、現在のキーワード検索サイト、特に、Google、Yahoo!、MSN など大手のサービスや機能の充実ぶりにはめざましいものがあります。検索サイトは、インターネットユーザーのポータルサイトとなることを目指して、激しい開発競争を行っています。検索エンジンそのものの機能だけで巨大なネットビジネスを作り上げた Google のように、検索サイトはビッグビジネスの糸口です。従って、開発は一層進み、検索機能、サービスがますます充実することは間違いありません。Google をはじめとする大手検索サイトは検索機能だけでなく、さまざまな Web サービスの提供を通じて、ポータルサイトとしての機能を充実させています。Google を例にすると 5GB に及ぶ巨大なオンラインストレージが利用できる G-mail やオフィスなみの機能を持つオンラインアプリケーション、写真や動画の共有機能などを無償で提供しています。現状では世界的に Google のシェアが突出した状況にありますが、あまりに急激に巨大化した Google に対する不安を抱くユーザーもあり、今後他の検索サイトがどのようなサービスや機能を提供してくるか、ビジネスとしての「検索」にも目が離せません。

■おさらい

- ・ 検索エンジンは事前に Web ページを収集している
- ・ 検索エンジンにはロボット型、ディレクトリ型、ハイブリッド型がある
- ・ 検索結果は検索エンジン独自のフィルタにより重み付けされ、表示順が調整される
- ・ キーワードの複数の組み合わせに記号を付けると、検索モードが切り替わる
- ・ 特定のサイトだけ、特定のファイル形式だけの検索もできる
- ・ すでに消えてしまった Web ページも探せばあるかもしれない
- ・ ジャンルごとに強みを持つ検索サイトがある
- ・ ブログは検索されやすい

参考

富士通 AzbyClub インターネット検索の達人ワザ

http://azby.fmworld.net/usage/vfami2/038/1_1/index.html

検索力トレーニング

<http://guide.search.goo.ne.jp/training/>

Google

<http://www.google.co.jp/>

Yahoo!

<http://www.yahoo.co.jp/>

MSN サーチ

<http://jp.msn.com/>