1 検索エンジンの概要

山名 早人（早稲田大学）
yamana@waseda.jp

村田 剛志（東京工業大学）
murata@cs.titech.ac.jp

検索エンジンの一般的な仕組み

検索エンジンは，一般的に**図-1** に示すように，大きく分けて，クローラ部，インデクサ部，検索部から構成される．クローラ部により，インターネット上から自動的に Web 情報を収集し，インデクサ部で検索のためのインデックスを作成する．ここで作成されたインデックスは，検索部で用いられる．

クローラ部

クローラは別名 Web ロボットやスパイダーとも呼ばれている．一般的なクローラは，**図-2** に示すように，起点として設定された Web ページを収集し（図中①），収集したページが持つリンクを解析し（②），未収集の URL を収集待ち URL キューに入れる（③）．収集待ち URL キューでは，優先度による収集順の入れ替えを

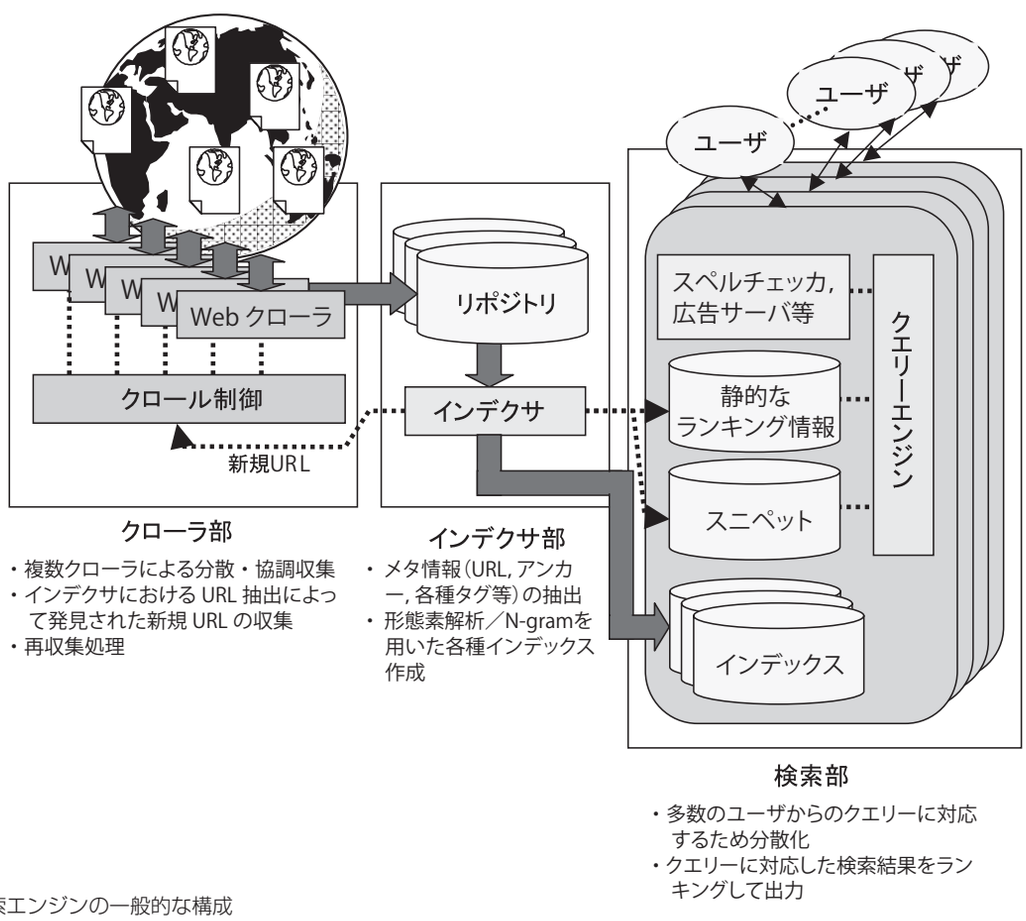

図-1　検索エンジンの一般的な構成

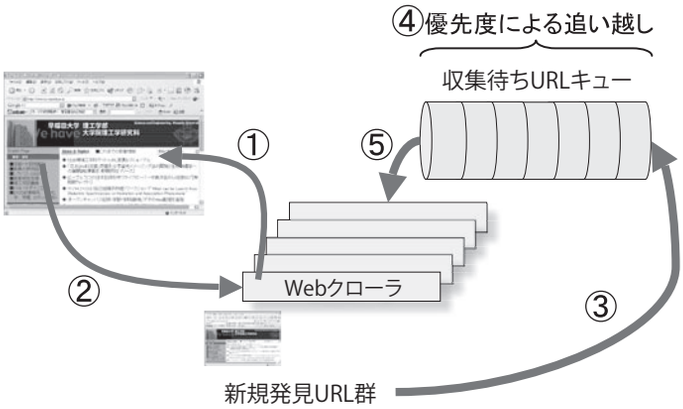

図-2　クローラの仕組み

	2週間後	1カ月後	3カ月後	6カ月後
更新ページの割合	4%	8%	15%	21%
トップページを更新したサーバの割合	20%	35%	52%	66%
トップページを更新したサーバ中の更新ページが全更新ページに占める割合	63%	76%	84%	88%

表-1　Webページの更新傾向[6)]

行い（④），優先度の高い URL から順に再びクローラで収集する（⑤）．そして，収集を高速化するために，複数のクローラを同時に動作させている．たとえば，IBM アルマデン研究所が実施している WebFountain プロジェクト☆1 では，48 台の 2.4GHz Intel Xeon 2CPU サーバ（インターネットとの接続は 75Mbps）を用いて Web ページの収集（5,000 万ページ／日）を行っている [1)]．

図 -2 の優先度キューにおける優先度の決定方法としては，さまざまな方法が提案されている．たとえば，index.html や welcome.html などで終わる URL は，一般的に，ある Web ページ群の中心となるページであると考えられるので，優先して収集する．また，アンカータグに記述された内容から当該 Web ページの重要度を判定したり，すでに収集済みの Web ページからの被リンク数に応じて重要度を設定したりする．文献 2）では，すでに収集済みの Web ページを解析し PageRank[3), 4)] を求め，PageRank が大きいページから優先して収集することにより，利用者がよく検索する重要なページを効率的に収集できることが示されている．

さらに，一度収集した Web ページを効率よく更新するために，一般的なクローラでは，Web ページの更新頻度に基づいて収集頻度を自動調整する．たとえば，収集間隔がデフォルトで 2 週間に設定されている場合を考える．今日，ある Web ページを収集したとする．次に同じページを収集するのは，デフォルトの 2 週間後である．そして，2 週間後に，（1）当該ページが更新されていれば次は 1 週間後に収集し，（2）更新されていなければ次は 4 週間後に収集するといった具合に，各 Web ページの更新頻度に応じて収集間隔を変える．

再収集アルゴリズムの中で最も効率のよい方法としては，スタンフォード大学の Cho☆2 らによる Web サーバごとの更新傾向を利用する手法がある．文献 5）では，Web サーバごとに Web サーバ内の Web ページの更新頻度が大きく異なることを利用して，収集対象を限定する手法が提案されており，著者の知る限り現時点で最も再収集効率がよい手法である．具体的には，各 Web サーバから一定数の Web ページを収集して，Web サーバごとの Web ページ更新割合を求め，更新割合の多い Web サーバから優先的に収集を行う．本手法は，著者らの調査 [6)] でも，その有効性が確認されている．**表 -1** は，2003 年 6 ～ 12 月の 6 カ月に渡り JP ドメイン内から任意に抽出した約 17,000 台の Web サーバを対象に，Web ページの更新傾向を調査した結果である．表 -1 に示すように，全体的に時間の経過とともに更新ページの割合が増加する．特に注目すべきは，「トップページ（/index.htm 等）が更新されている Web サーバ（全体の 20%）のみを収集対象とするだけで全更新ページの 63%を収集できる（表 -1 の 2 週間後のカラム）」という点である．

これまで述べたような Web ページごと，あるいは，Web サーバごとに更新頻度を変えることなく，毎回 100 億にも及ぶ Web ページを収集すると仮定すると，Web ページの平均容量を 20KB，使用できるネットワークの平均使用帯域幅を 100Mbps としても，すべてのページを収集するのに半年以上が必要となる．つまり，半年より短い周期で更新される Web ページをうまくインデックス化することができない．

以上述べた以外にも，最近の商用検索エンジンでは，全 Web 空間を対象とした一般的な Web クローラとは別に，ニュースやブログのように頻繁に更新されるページのみを収集対象とした特別なクローラが用いられている．

☆1　WebFountain は，検索エンジンに対して一般的に「分析エンジン（Analysis Engine）」と呼ばれるものであり，Web ページを含む膨大な情報の中から有用な情報を見つけ出すことを目的としたエンジン．

☆2　現在は，カリフォルニア大学ロサンゼルス校の助教授．Web ページの再収集に関する世界的な第一人者．

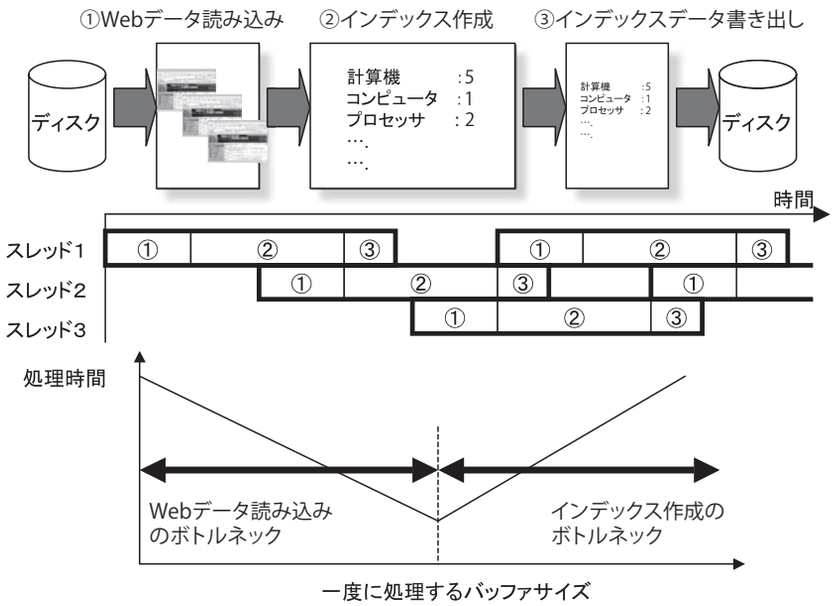

図-3　インデックス作成の高速化

このようなクローラは 1 日に何回も同じページを収集しインデックス化することによって，最新の情報も検索できるようにしている .

インデクサ部

Web 情報のインデキシングには膨大な時間が必要となる．これは，収集された Web ページから形態素解析による自動キーワード抽出や，キーワードを付与せずに N-gram 文字インデックスを作成するための CPU 処理時間とディスクアクセス時間が膨大となるためである．

これを解決するために，スタンフォード大学の Melnik らは，文献 7）においてマルチスレッドを使用してソフトウェアパイプライニングを実現する手法を提案している．

図 -3 に示すように，インデキシングの行程を① Web データ読み込み，②インデックス作成，③インデックスデータ書き出しの 3 つの行程に分割し，これらの 3 行程が並列に動作するように，3 つのスレッドによりソフトウェアパイプライニングを実現する．これにより，あるスレッドがディスク I/O 待ちになっている間も CPU を有効利用することができ，高速化が可能となる．図 -3 に示すように，一度に処理するバッファサイズ（データ量）を増減させることで，①の Web データ読み込みがボトルネックとなったり，②のインデックス作成がボトルネックになったりするので，処理時間が最小となるようにバッファサイズを調整する．

一般の商用検索エンジンにおけるインデキシングの詳細は公開されていないが，Google では，2003 年 6 月までの運営において，Web 情報収集に約 1 カ月，インデックス作成に約 2 カ月を費やしていた [4]．そして，この収集・インデキシングという一連の作業を行うコンピュータ群を 3 セット用意することにより，あるコンピュータ群がクロールしている最中に，別のコンピュータ群では，それぞれ 1 カ月前，2 カ月前に収集した Web 情報のインデキシングを行う．これにより，約 3 カ月前の Web 情報が毎月インデックスとして更新されることになる☆3．

検索部

検索部は，図 -1 に示すように，インデクサにより作成されたインデックスと，検索結果とともに出力するスニペット☆4，PageRank に代表される静的なランキング情報，検索結果に付加価値を付けるためのスペルチェッカや広告サーバから構成される．また，ユーザからのリクエストに対するターンアラウンドタイムを短縮するために，検索部自体を分散化するのが一般的となっている．

たとえば，Google では 1 日に 2 億件☆5 を超える検索を処理するために，世界中のデータセンタに負荷を分散させている．これには，ダイナミック DNS が用いられる [8]．

通常，DNS（Domain Name Server）の TTL（Time To Live）は 12 ～ 24 時間に設定されるが，Google では，

☆3　現在の Google では，更新頻度の高い Web ページに対して専用クローラを用いることで，毎日収集・インデックス更新を行っていると推測される．
☆4　検索結果として出力される各 Web ページのタイトル下に表示される Web ページの内容を表した数行の文．
☆5　2003 年 8 月時点のクエリー数．

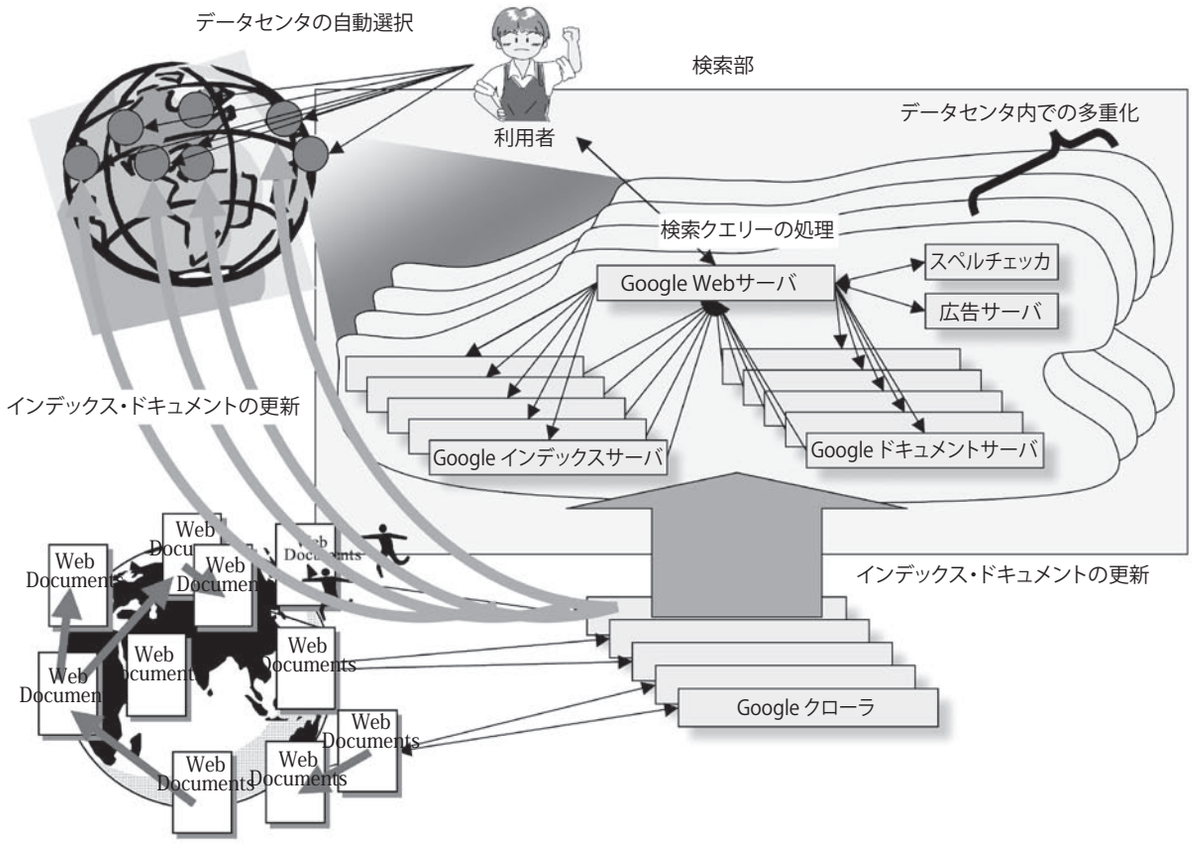

図-4　Googleの全体構成（文献8）を基に作成）

これを5分に設定している．TTLは，DNS情報をキャッシュしているDNSサーバが再びオリジナルのDNS情報を参照するまでの間隔である．これが5分に設定されていると，5分後には必ずGoogleの新しいDNS情報を参照することになる．この仕組みを使って，一番負荷が軽いGoogle WebサーバのIPアドレスが指定され，負荷分散が実現される．なお，Googleはこの負荷分散をデータセンタ間の負荷分散と，データセンタ内での負荷分散の2段階に分けて行っている．

図-4にGoogleの全体構成と検索部の詳細を示す．利用者がGoogleにアクセスすると，世界中の10カ所を超えるデータセンタの1つが自動的に選択され，選択されたデータセンタ内の1つのGoogle Webサーバが応答する．Google Webサーバは，検索クエリーを受け取るとすぐに複数のGoogleインデックスサーバに同時に検索要求を送る．そして，返ってきたインデックス情報に基づいてGoogleドキュメントサーバに指示を出し，検索クエリーに応じた検索結果のページを作成する．

Googleドキュメントサーバには，生のWeb情報（Webページ，画像，PDF，Usenetニュース等）が保存されている．Webページを対象とした検索では，まず，Googleインデックスサーバで検索されたWebページのIDが，Google Webサーバを経由してGoogleドキュメントサーバに送られる．すると，Googleドキュメントサーバにおいて，該当するWebページのタイトルや検索クエリーで指定されたキーワードと一致する部分を抜粋したスニペットが作成され，Google Webサーバに送り返される．

また，検索クエリーとして入力された語にスペルミスがないかを確認したり，日本語の場合には「コンピュータ」や「コンピューター」のような表記のゆれに自動的に対応し，クエリー拡張を行うのがスペルチェッカの役目である．クエリー拡張とは，たとえば検索クエリーとして「コンピュータ」が入力された際に，「コンピュータ OR コンピューター」のように検索クエリーを自動的に拡張する仕組みのことである．さらに，広告サーバは，検索クエリーに対応する広告を抽出する役目を果たす．

主要な検索エンジン

検索エンジンの日本国内でのページビュー☆6は，ネットレイティングス社の2004年12月の1カ月間の調査によると，Yahoo!が約16億ビュー，Googleが約10億ビュー，MSNが約4億ビューとなっている[9]．第4位以下は1億ビュー未満であり，Yahoo!，Google，MSNが三大検索エンジンとなっていると言える．

☆6　利用回数．

検索エンジン	インデックス規模	検索対象言語	検索対象ファイル形式	検索オプション（AND, OR, NOT, フレーズ検索を除く）	その他の特徴
Yahoo! http://search.yahoo.com/	42億	37	7	ドメイン指定検索 言語指定検索 国指定検索 ファイル形式指定検索 日付範囲指定 アダルトコンテンツ除外指定検索	自動的に検索語を抽出するY!Q検索をサポート http://yq.search.yahoo.com/
Google http://www.google.com/	81億	105	7	ドメイン指定検索 言語指定検索 ファイル形式指定検索 逆リンク指定検索 タイトル限定検索 URL限定検索 本文限定検索 日付範囲指定検索 アダルトコンテンツ除外指定	
MSN http://search.msn.com/	50億	12	不明	ドメイン指定検索 国指定検索 逆リンク検索	ランキング時のパラメータとして，当該ページの更新頻度，人気度，および，表記ゆれの度合いを指定可
Teoma http://www.teoma.com/	20億	10 （日本語無）	不明	ドメイン指定検索 言語指定検索 地域指定検索（アフリカ，中央アメリカ等） タイトル限定検索 URL限定検索 日付範囲指定検索	検索結果の自動クラスタリングによる絞り込み検索語提示機能

表-2　主要検索エンジンの比較

そのほか，特色を持った検索エンジンとして，検索結果の絞り込みのためのキーワード表示機能や検索結果の保存・管理機能を備えた Ask Jeeves☆7，検索結果を自動的にクラスタリングしてくれる Mooter☆8 などがある．特に Ask Jeeves はバックエンドに Teoma☆9 を使い，検索結果のランキング計算に，Subject-Specific Popularity☆10 と呼ぶ指標を用いている．Subject-Specific Popularity とは，検索対象となる主題に合致する Web ページ群内で Authority☆11 となるページに高い重要度を与えランキングする方法である．Google がランキングに使用している PageRank は，すべての Web ページをもとに計算されるのに対して，Subject-Specific Popularity は，同一の主題を持つページ内で計算されるという点に違いがある☆12．簡単に言えば，自動車に関する検索をした場合に，自動車に関係する Web ページ集合内での被リンク関係のみを用いてランキング計算をする．これによって，たとえば，花屋さんのページからのリンクなど，主題以外のページがランキングに及ぼす影響を排除しランキング精度を上げることを目指している．また，Teoma は，検索結果を自動的にクラスタリングする機能を持ち，たとえば「Apple」で検索した場合，「Apple Juice」「Apple Computer」「Apple Mac OS」などのようなクラスを表示してくれる．**表 -2** にこれまで述べた主要な検索エンジンの機能比較を示す．

検索エンジン間の関係は，提携，買収，再編などによってめまぐるしく変化している．2005 年 6 月現在における検索エンジン間の提携関係を表す相関図を**図 -5** に示す．先に述べたように，主に Google，Yahoo!，MSN の 3 つのグループから構成されていることが分かる．

検索エンジンの提携関係が変化する要因としては，新たな技術の台頭や企業戦略の変化などが挙げられる．たとえば Yahoo! JAPAN は 1999 年には goo と提携して検索サービスを行っていたが，2001 年には Google を採用している．さらに 2004 年にはその提携を解消して自社製の検索エンジンを使用している．また提携においても，提携先の検索サービスを全面的に用いるのではなく，自社の独自技術を併用する場合もある．たとえば goo は，入力キーワードの日本語処理を行った上で Google での検索を利用しており，Google 単独の場合とは異なる検索結果を得ることができる．

☆7 http://ask.com/，日本語版は http://ask.jp/
☆8 http://www.mooter.com/
☆9 http://www.teoma.com/
☆10 http://sp.teoma.com/docs/teoma/about/searchwithauthority.html
☆11 被リンク数の多いページ．
☆12 同一主題のページをどのように求めているかについては，公開されていない．

検索エンジンにまつわる話題

検索エンジンの歴史

検索エンジンの歴史は非常に新しい．米国で Yahoo! が設立されたのは 1995 年，Google が設立されたのは 1998 年である．現在までの 10 年程度の検索エンジンの歴史を振り返ると，乱立・淘汰・新たな展開の 3 つ

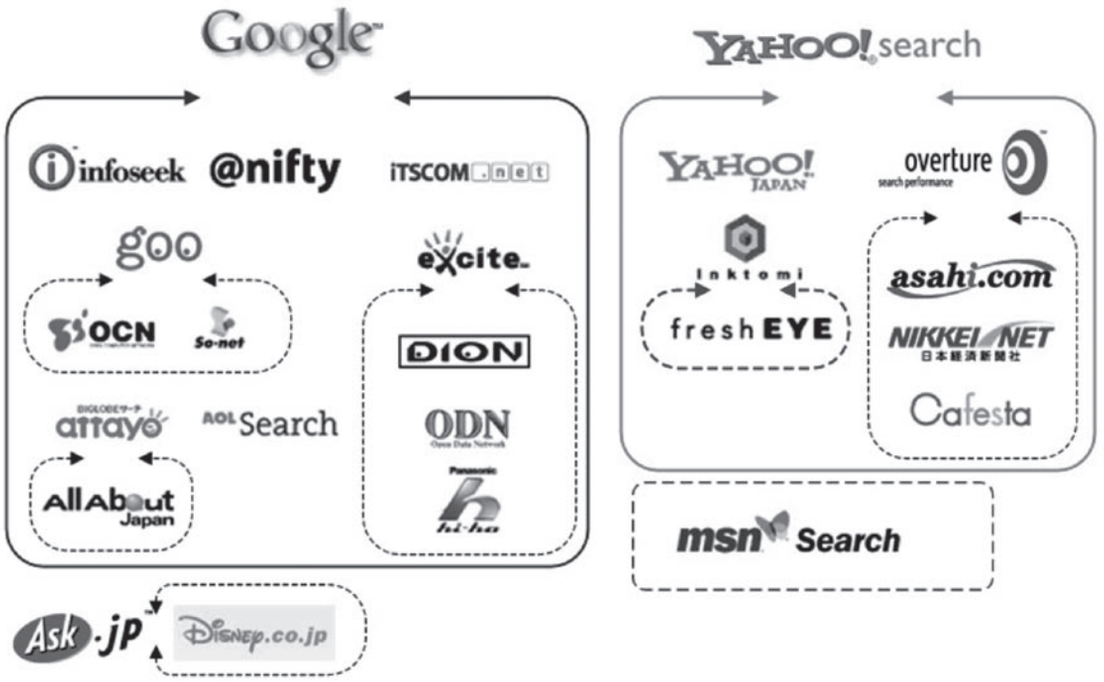

図-5　検索エンジンの相関図（サーチエンジンマーケティング総合研究所 渡辺隆広様より提供）

のフェーズとして捉えることができる．1990年代後半においては，アクセス数を増やすことがいずれは収益を生むという発想のもとに，Web利用者が最初に訪問するポータルサイトとして多くの検索エンジンがサービスを開始した．その後，ポータルサイトが持つ多様な情報コンテンツの中の一機能へと検索サービスの位置づけが変化すると，検索エンジンを自社で開発・運営するコストを敬遠し，主要な検索エンジンと提携する動きが盛んになってくる．2000年代初頭においては，IT不況の影響による検索サービスの消滅や，買収や統合などによる再編などによって淘汰が進んでいる．たとえば楽天は2000年にインフォシークジャパンを買収，2002年にライコスジャパンを完全子会社化し，現在では同社のポータルサイト事業として運営されている．そしてWebが社会インフラとして普及した現在においては，多くの利用者の目にとどまる広告媒体として検索エンジンが再び注目されてきている．

キーワード広告

Web上の広告としては，サイト上の一定スペースを占有して宣伝するバナー広告のほかに，ユーザが入力したキーワードの検索に関連するサイトを検索結果とともに提示するキーワード広告がある．後者は検索エンジン特有のものであり，キーワードを入力したユーザのニーズに合った広告が期待できる．検索結果への広告の付加は，古くはAltaVistaによって1999年に試みられたものである．広告枠をオークションで売り，クリックされた分だけ広告料を課金するというスポンサードサーチのモデルは，現在では数十億ドルのビジネスへと成長しており，検索エンジンの収益の柱の1つとなっている．

オーバーチュアやGoogleによるスポンサードサーチは，ユーザが入力した検索キーワードに関連した広告を検索結果とともに提示する．各広告にはクリック単価が設定されており，ユーザがその広告をクリックした回数を掛けた広告料が広告主に課金される．広告主は，自社の広告を出すキーワードと，そのクリック単価を入札することで広告出稿を行う．同一のキーワードに複数の広告主が出稿した場合，入札された額に応じて，オークションによってクリック単価および広告の掲載順位が決定する．オーバーチュアの場合は入札された額の高い順に広告が表示され，Googleの場合は広告のクリック率も加味して広告が順位づけされている．

クリック詐欺

このような広告モデルに対して，ライバル会社の広告のクリック数を意図的に増やし，それによって余計な広告費をライバル会社に使わせてしまうことを目的とした，クリック詐欺と呼ばれる現象が起こっている．具体的には，自動的にクリックするプログラムを使ったり，安い労働力を使ってひたすら手作業でクリックするなどの手口が使われている．検索エンジン側がクリック詐欺に対して公式にコメントすることはほとんどないが，スポンサードサーチが注目を集めていることの負の側面であると言える．

検索エンジン最適化（SEO）

企業が自社および自社製品を Web で宣伝するにあたり，関連するキーワードで検索した際に，検索結果の上位に自社サイトを表示させることが必要になってきている．Web 利用者の多くは，「検索結果の上位に表示される企業はメジャーブランドである」と考える傾向にあり，検索結果の上位のサイトしか見ないためである．このことから，検索結果表示で自社サイトを上位に表示させることを目的とした Web ページの改良が行われており，検索エンジン最適化（SEO）と呼ばれている．これは検索エンジンのページ収集や検索結果のランキングを推測し，ページ上で使用する単語を変更したり，他ページとのハイパーリンクによる結合を強化したりすることによって行われている．検索結果のランキングに対するこのような人工的な操作は検索エンジン側にとっては望ましいものではなく，しばしば SEO と検索エンジンとのいたちごっこなっている．

検索エンジンの実験的な試み

ユーザの情報ニーズに応える検索を実現するために，数多くの新たな試みが現在も進行中である．実験的な検索サービスを公開している例として Google Lab（http://labs.google.com/）や goo ラボ（http://labs.goo.ne.jp/）が挙げられる．Google Lab においては，本格的な検索サービスに取り入れられたものとして，後述するデスクトップ検索や，単語の定義を出力する Google glossary などがある．goo ラボについては，本特集の Goo の記事中で紹介されている．

Google Web API

検索エンジンは通常，ユーザがキーワードを入力して検索を行うが，コンピュータプログラム上から検索エンジン資源を利用することによって豊富な Web 情報を活用したプログラムの実現が期待できる．そのような試みとして，Google が 2002 年に公開した Google API がある．これは，SOAP および WSDL に基づいたインタフェースで，Google が収集した Web ページのデータを自分の好きなようにプログラミングして利用することができるものである．多くの利用者がその活用を試みているが，面白い例としては Microsoft Word に Google の機能を加えることで新しい流行語などのスペルチェックを可能にした CapeSpeller（http://www.capescience.com/google/spell.shtml）や，Google の検索結果のグラフィカルな視覚化を行う TouchGraph Google Browser（http://www.touchgraph.com/TGGoogleBrowser.html）などがある．また Yahoo! も 2005 年に Web API を公開している．

デスクトップ検索

デスクトップ検索は，自分のパソコン上にあるファイル等を検索するものであるが，従来から OS に備わっていた全文検索機能とは異なり，検索対象のテキスト情報をすべて事前にデータベース化しておき検索を行う．2004 年末から 2005 年にかけて，Google，Microsoft，Yahoo!，Ask Jeeves などが相次いでデスクトップ検索ツールを公開している．ユーザ側のニーズとして，個人のパソコン上に蓄えられている膨大なファイルを検索したいことはもちろんである．それと同時に検索エンジン側も，検索精度の向上やパーソナライズ等，将来の検索サービス機能の充実を行うための情報源として，デスクトップ検索に注目していると考えられる．

まとめ

本稿では，検索エンジンの一般的な仕組みについて説明するとともに，主要な検索エンジン例とその提携関係，さらに検索エンジンをめぐる最近の話題を紹介した．ネットレイティングス社の調査によると，Google を使用しているユーザの 6 割が他の検索エンジンも併用している．これは，多くのユーザが単一の検索エンジンの結果に十分に満足していない現状を表しているとも言える．検索エンジンはまだまだ発展途上であり，新たな個性的な検索エンジンの出現によって勢力地図が大幅に塗り変えられる可能性は十分にある．現状では主要な検索エンジンのシェア争いが注目されがちであるが，各々の検索結果の特徴を正しく理解することがユーザに求められていると言える．

参考文献

1) Cass, S.: Fountain of Knowledge, IEEE Spectrum, Vol.41, No.1, pp.68-75 (2004).
2) Arasu, A., Cho, J., Molia, H.G., Paepcke, A. and Raghavan, S.: Searching the Web, ACM Trans. on Internet Tech., Vol.1, No.1, pp.2-43 (2001).
3) Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Proc. of the 7th WWW Conf., pp.161-172 (1998).
4) Calishain, T. and Dornfest, R. (田中訳, 山名監訳): Google Hacks, オライリー・ジャパン (2003).
5) Cho, J. and Ntoulas, A.: Effective Change Detection Using Sampling, Proc. of the 28th Int. Conf. on Very Large Databases, pp.514-525 (2002).
6) 熊谷, 山名：リンク構造を利用した Web ページの更新判別手法, DEWS2004 予稿集, 5-B-02 (2004).
7) Melnik, S., Raghavan, S., Yang, B. and Molina, H. G.: Building a Distributed Full-Text Index for the Web, Proc. of the 10th WWW Conf. pp.396-406 (2001).
8) Barroso, L. A., Dean, J. and Holzle, U.: Web Search for a Planet: the Google Cluster Architecture, IEEE Micro, Vol.23, No.2, pp.22-28 (2003).
9) 検索上手になる 5 つの「ツボ」：欲しい情報が短時間で手に入る, 日経パソコン 2005 年 2 月 28 日号, pp.72-95 (Feb. 2005).
10) 神崎洋治, 西井美鷹：体系的に学ぶ検索エンジンのしくみ, 日経 BP ソフトプレス (2004).

（平成 17 年 6 月 29 日受付）