情報検索システムとしてみたサーチエンジン

久野 高志（作新学院大学女子短期大学部）
安形 輝　（亜細亜大学国際関係学部）
上田 修一（慶應義塾大学文学部）

【抄録】Web サーチエンジンは 1994 年頃から公開されはじめたが，わずか５，６年の間に急速な発展を遂げている。情報検索分野で研究されてきた検索技術は，大規模で不定形でありハイパーリンクを持つ Web ページの検索に利用されることになった。まず，サーチエンジンの発展過程を示し，その中で利用されている手法やアイデアを分析し，従来の情報検索の手法と照合することにより，情報検索におけるサーチエンジンの位置付けを明らかにする。

１　サーチエンジン出現の背景

1990 年代初期の商用化されたばかりのインターネットの世界は混沌としており，最初に注目されたのは Gopher だった。Gopher はメニュー形式により，インターネット上に散在するファイルやデータベース，さらには Archie などのサーバーを階層的に結びつけた。

また，1993 年９月に公開された Web ブラウザーの Mosaic は，それまでラインモードでの閲覧が主流であった WWW のインタフェースを大幅に改善した。これにより WWW は一気に普及することとなる。この時期，Gopher や Web ページのもつリンク機能は参考図書の索引や図書目録などに例えられ，Gopher はメニュー形式，Web ページはマルチメディア対応の「情報検索システム」とされていた[1]。

こうした点から，当初はリンク機能を多用したリンク集的サイトがインターネット上での探索ツールとして位置付けられていたことがうかがえる。例えば，わが国においては「NTT Home Page」が 1993 年 10 月に公開され，その先駆け的役割を担っていた。

その後，WWW の急速な普及にともない，リンク集の主流は Web ページ上のものへと移行していき，さらにリンクを独自のカテゴリーにより分類するようになった。1995 年 4 月にサービスを開始した Yahoo!(http://www.yahoo.com/)は，新しく作られた Web ページを主として登録制で収集し列挙型分類を用いて主題からの階層的な探索ができるようにした。このほか，「The WWW Virtual Library」，「The Whole Internet Catalog」，「EINet Galaxy」などはいずれも階層構造をもったカテゴリーリストを備えていた。初期のサーチエンジンは，こうした分類を用いたものが主流だった。

一方，ロボットを用いたサーチエンジンは，1994 年頃より出現した。「Infoseek」，「Lycos」，「WebCrawler」などが初期の代表的なものである。これらはディレクトリ型と比較した場合，収録データが大規模で Web ページのほぼ全体を対象に全文検索ができるという利点をもっていた。しかし，データそのものが未整理の状態にあり，検索結果が膨大になりがちであるという問題などが指摘された[2]。

２　サーチエンジンで応用されている技術

サーチエンジンのこれまでの発展経緯をたどると，従来の情報検索分野において研究されてきた検索技術が初期の段階から用いられていることがわかる（表１参照）。これは当初から，従来の検索システムの延長上にサーチエンジンが位置付けられていたためだと思われる。

しかしながらその後の展開を追っていくと，WWW の特性と情報検索手法とを関連付けしつつサーチエンジン特有の検索手法の開発が行われてきていることがわかる。

２．１　従来の情報検索を応用した技術

・論理演算子／近接演算子

Web ページの検索は，一般的な検索語では数多くの検索結果が出力される傾向にある全文検索である。そのため，どのサーチエンジンでも複数の語を使い検索式を構築することができる機能が用意されている。ただし，フレッシュアイ（http://www.fresheye.com/）の調査によれば，約 8 割の利用者はキーワードを 1 語のみ用い，約 3 割は検索結果を絞り込むことができないなどの理由で 4 回に 1 回は検索をあきらめている[3]。そこで，ほとんどのサーチエンジンでは複数の語を演算子なしに入力すると標準で AND 検索を行う。

・順位付け出力

検索文献を適合度順に順位付け出力することは，従来，研究においては G. Salton の SMART システムをはじめ数多くのシステムで応用されてきた。しかし，商用の情報検索システムでは DIALOG のコマンドの一つである TARGET のような形でオプションとして用意される程度であり，本格的に利用されてきたとは言いがたい。しかし，Web ページの検索は①全文検索，②膨大なページ数，という二つの理由から，検索結果が数多く出力されてしまうため，順位付け出力はロボット型サーチエンジンでは登場直後から実現されてきた。

適合度順出力における文献の重み付け手法は当初，語の重み付けを用いていたが，Google(http://www.google.com)の成功以後は何らかの形でリンク関係の分析を応用することが多くなっている。

・適合性フィードバック

goo(http://www.goo.ne.jp)に代表されるいくつかのサーチエンジンでは適合性フィードバック

を応用し，サービス当初より検索結果から複数のページを選択し，それを次回の検索のキーとすることができた。しかし，チェックボックスというインタフェースはわかりにくく多くの利用があったとはいえない。現在も適合性フィードバックを行っている Google では複数ページをキーとしたフィードバックは行わず，単に「関連ページ」というリンクによって実現している。

・検索式の拡張

検索式中に含まれる語と類似した検索語を提示する，あるいは，自動的に検索式中に含めることである。自動的に検索式を拡張するシステムは少なく，「関連語表示」という形で実現しているものが多い。サーチエンジンで検索式の自動的拡張が採用されていない理由は，これによって検索語が含まれないページが出力されることがあり，一般的な利用者が混乱する恐れがあるためと考えられる。

・あいまい検索／自然言語処理

あいまい検索，ファジー検索と呼ばれるアルファベットの大文字・小文字，日本語の半角・全角の区別をしないなどの手法は，利便性のため情報検索で従来から行われてきたことであり，サーチエンジンでも最初から行われてきたことである。

自然言語処理技術を使用していることを明示しているサーチエンジンは少ないが，日本語の特性からどのサーチエンジンも多少の差はあっても応用していると考えられる。

・フィールド指定

Web ページは一般的なデータベースのように明確な構造化がされてはいないが，タグ情報を使うことである程度構造的に分析することができる。その情報をほとんどのサーチエンジンではメタタグのキーワード情報やタイトルタグの重み付けを上げる形で暗黙のうちに利用するが，一部のサーチエンジンでは明示的に指定できるものもある。また，リンクタグで囲まれた説明語句から検索できるものもある。

・多言語検索

1995 年に公開された「NIPPON SEARCH ENGINE」は，和英･英和の翻訳とローマ字によるキーワード入力機能を備えていた。このように早い時期から多言語への対応は検討されている。現在では，翻訳よりも検索対象としての言語を選択することにより対応するものが多い。また，国ごとに独立したサーチエンジンを運営し相互に

表1　サーチエンジンで使われている手法

情報検索／インターネット	手法／技術	サーチエンジンでの用語	手法／技術の説明	代表的なエンジン *は最初に応用
情報検索	論理演算子	検索条件の指定(goo)	複数の検索語関係を論理演算子を使い表現する	excite, Lycos
情報検索	フレーズ検索／近接演算子	フレーズ(goo) NEAR検索	･熟語をそのままの形で検索する ･複数の語の出現位置を指定する	Lycos
情報検索	順位付け出力	スコア順表示(infoseek)	検索結果を適合度順に出力する	Google, goo
情報検索	あいまい検索		大文字小文字、全角半角、スペルミスなどを同一視して検索を行う	goo, infoseek
情報検索	自然言語処理	自然言語認識(infoseek)	自然言語の形で入力された検索式から検索を行う	infoseek(*)
情報検索	フィールド指定	ページのタイトル(lycos)	Webページ中の指定部分からの検索を行う	Lycos
情報検索	適合フィードバック	関連ページ(Google)	検索結果中の文献をキーとして再検索を行う	Google, RCCAU Mo-n-do-u(*)
情報検索	検索式の拡張	つぼシーク(infoseek)	検索式と関連が高いと思われる語を提示する	AltaVista, infoseek
情報検索	多言語検索	翻訳検索(excite)	･検索式の言語とは異なる言語のページを検索する ･検索結果を他の言語に変換して表示する	excite, TITAN
情報検索	マルチメディア検索	画像･音声検索	画像や音声ファイルを対象として検索を行う	Google image search
情報検索	データベースの指定	厳選サイトのみ(lycos)	分野、有用性などから制限されたデータベースを対象とした検索を行う	Lycos, Northern Light
情報検索	自動分類	自動分類(TITAN)	･収集したWebページをカテゴライゼーションする ･検索結果を自動的に分類する	TITAN, Northern Light
情報検索／インターネット	引用検索	リンク検索(infoseek)	リンクしている／されているページを検索する	goo, infoseek
情報検索／インターネット	引用分析	PageRank(Google)	リンク関係の分析により順位付け出力を行う	Google(*), goo
インターネット	URL検索	サイト検索(infoseek),ドメイン検索(Lycos)	URL中の語句から検索を行う	infoseek, goo
インターネット	ページ群の集約	サイトごとの表示	同じサイト内のページをまとめて表示する	Google, goo
インターネット	サムネイル表示	リンク先画像	検索結果のページの縮小表示を行う	ODIN(*), goo
インターネット	更新ページの収集	新鮮情報（フレッシュアイ）	ロボットによる収集方法を工夫することで新しいページを検索する	フレッシュアイ(*), Google
インターネット	ページタイプの判定	ページタイプサーチ(Netplaza)	「リンク集」といったページタイプ別に検索ができる	Netplaza

リンクをかけている場合もある。

・マルチメディア検索

画像，音声などを対象とした情報検索は以前から研究されてきたが，標準的なアルゴリズムが確立されているとはいえない。サーチエンジンにおける応用も画像のファイル名やリンクからの検索に留まっている。

・データベースの指定

一般的な Web ページはあまりにも膨大であるため，いくつかのサーチエンジンでは収集対象を諸機関の公的ページや有料の情報源などに限定したデータベースを一般的なデータベースとは別に，あるいは組み合わせて用意することで，より精度の高い検索を提供している。これは従来の商用データベース検索において専門分野ごとにデータベースを構築してきたことに近いといえる。

・自動分類

文献の自動分類は，クラスタリングとカテゴライゼーションに大きく分かれるが，すべてのデータを対象にするクラスタリングを膨大な量が存在する Web ページに対して行うことは現実的ではない。そのため，全体的なデータに対する自動分類はカテゴライゼーションの形で行われている。また，Northern Light などのサーチエンジンでは検索結果に対する分類を行っているが，技術的には関連語の表示に近いと考えられる。

２．２　情報検索とインターネットの両方に関わる技術

・引用検索

情報検索においては引用索引を用いた検索が従来から行われてきた。リンク関係を引用関係と見なすならば，サーチエンジンにおけるリンク検索は引用検索の応用と考えることができる。

・引用分析の応用

Google における PageRank アルゴリズムは，被リンク数が多いページや被リンク数が少なくても被リンク数が多いページからリンクをされているページを重要なページと見なし，検索結果の上位に出力するものである。これもリンク関係を引用関係と見なせば，引用分析の応用と考えることができる。

２．３　インターネット特有の技術

・URL 検索

サーチエンジンによってはページの URL 中の文字列から検索を行うことができるものがある。URL からの検索は技術的には高度なものではなく単なる文字列の一致である。しかし，URL を構成するドメイン名，パス名，ファイル名に重要な情報が含まれる場合があり，インターネット上では有効な検索手段の一つになりうる。

・ページ群の集約／サイトごとの表示

一般的な情報検索では文献単位で検索を行うが，サーチエンジンは Web ページを単位として行うため，同じサイト内のリンクで結ばれている類似したページが数多く出力されてしまうことがある。そのため，最近のサーチエンジンではサイトごとにまとめて表示することが一般的になりつつある。

・サムネイル表示

従来のサーチエンジンの多くが，テキストでリンク先のページの概要を表示し，リンク先ページが必要とするページかの判断がある程度可能なようにしてきた。しかし，Web ページが有用かの判断において，レイアウト情報も判別できるよう縮小した画像を表示するサーチエンジンも登場している。技術的には高度であるが，有効であるかは不明である。

・更新ページの収集

単純なロボットを使い収集されたページを更新するには，Web ページの膨大な量とネットワークの物理的な制約から数週間以上かかるが，更新頻度の高いページを中心として収集を行う工夫を行うことで，数日以内の新着あるいは更新ページを検索することができるサーチエンジンがある。

・ページタイプ判定

NEC の Netplaza では，Web ページの主として形態的な特徴に基づき，いくつかのページタイプを設定し，「リンク集を対象とする」などの検索を用意していた。

３　サーチエンジンに影響する技術以外の要素

サーチエンジンを商用ベースで運用していくには広告収入に頼ることになる。そのため，支配的なシェアを取ることが収入に直結するため，以下のような手法をとることがある。

・デフォルトサイト（ポータルサイト）

利用者がサーチエンジンを選択する際の基準は，機能的，収録データ的な点のみではない。例えば，MSN Search(http://search.msn.com/)は圧倒的なシェアを占める Internet Explorer の標準のホームページを MSN に，標準サーチエンジンを MSN Search にすることでシェアを急速に拡大していると思われる。

・多くの情報サービスの集約

多くのサーチエンジンでは一つのページに表示できるかぎり，数多くの情報源を提示し，利便性が高いことをアピールし，ポータルサイトとして使われることを目指してきた。

・ディレクトリ型とロボット型サーチエンジンの境界の曖昧化

ディレクトリ型とロボット型サーチエンジンはそれぞれ長所短所があり，それを補うためにお互いの機能を取り込んだり，提携をすすめてきた。goo や Google は圧倒的なシェアを占めるディレクトリ型サーチエンジンである Yahoo!と提携することで利用を増加させた。

・検索結果の優先順位

サーチエンジンにおいてバナー広告収入以外に収入を得るための手段として，検索結果の提示順の優先度を販売しているエンジンも登場している。

表２　手法の導入

	サーチエンジン	手法／技術
1995.11	NIPPON SEARCH ENGINE	多言語検索（翻訳，ローマ字）
1996.4	Infoseek WebCrawler	自然言語処理
1997.2	AltaVista	検索式の拡張（関連語の表示）
1997.8	Northern Light	検索結果の自動分類
1998.6	フレッシュアイ	更新ページの収集
1999.2	HotBot	検索結果中の利用者の選択をモニター，解析
1999.4	Lycos HotBot	ディレクトリ型検索サービス導入
1999.6	Netscape Search	人手で評価されたサイトのディレクトリ検索
1999.9	Google	ページランク

４　主要なサーチエンジンとその動向

サーチエンジンは，現在強い競争下におかれており，短期間に急速な進展をみせている。

もともと，キーワード検索時における各種演算子やトランケーション，フィールド限定，結果表示における出力数指定などの機能は備わっていた。その後，自然言語処理（Infoseek ほか，1996），検索式の拡張（AltaVista，1997），検索結果をカテゴライズして表示（Northern Light，1997），など様々な手法が導入された（表２参照）。同時に検索機能とは関係のない部分で，ニュース・天気，掲示板等のユーザーコミュニティ，ユーザーカスタマイズページや無料メールアドレスなどによるパーソナル化，といった機能の提供などが行われ，WWW への総合ポータル化がすすんだ。

そのような中1999年9月に公開されたGoogleは，シンプルなページレイアウトからなり，いわゆる総合ポータル的サーチエンジンではなく，公開当初から検索ポータルとしての色合いを強く持つサーチエンジンであった。しかし，２．２で述べた引用分析的アプローチによる独自のリンク関係分析アルゴリズムを実装し，2001 年 5 月時点には利用率で Yahoo!に次ぐ 2 位にはいっている（表３参照）。これは，PageRank により，多様な検索オプション選択の過程を経ずに Web ページの有用性がある程度自動的に検索結果に反映されたことが利用者に支持されたと思われる。同時に総合ポータルではなくても，検索目的に特化したサーチエンジンが利用者によって他のポータルと使い分けられているということも考えられる。また，特定の目的やトピックごとに検索対象を限定した「目的特化検索エンジン」が今後広く利用されるという見方がある [5]。

Webseed 社が 2001 年 1 月に発表した調査結果によると，PageRank のようにリンク構造をもとにした Web ページの「人気度」を検索技術として採り入れているサーチエンジンは有名サーチエンジンのうち 95%を占めるとされている [6]。

５　まとめ

以上からインターネットのサーチエンジンの発展について次のようなことがいえる。

①初期のサーチエンジンは情報検索分野で研究されてきた技術の応用を中心として発展してきた。

②WWW の特性が考慮されるようになった。特に 1999 年以降は，共引用とのアナロジーでリンク関係が検索に用いられはじめ，これは大きな転機となった。

③検索技術の複雑化の一方で，インタフェースは，多くの情報源を提供するポータル的な方向と極力簡素にする方向に分かれる

④ディレクトリ型のサーチエンジンは，依然として多くの利用者から利用されている。

検索技術的には上記②で述べたように，WWW の特性が考慮されながら発展を続けているが，様々な手法が試行錯誤的にテストされている段階であることにかわりはない。Google の成功でリンク構造の応用に関心が集まり多くの手法が試されている一方で，例えば大量のデータ収集機能を前提としたデータマイニングの応用などが検討・開発されてもいる。

また，「ボータイ理論」[7] などの研究によって指摘されてきたリンクにより辿ることができないページをどのように収集するのかなど，間接的課題も多い。

表３　サーチエンジンの利用率 [4]

	1999	(%)		2001	(%)
1	Yahoo!	45.92	1	Yahoo!	41.47
2	Excite	21.68	2	Google	13.87
3	AltaVista	9.70	3	MSN	12.91
4	Infoseek	5.32	4	AOL	5.40
5	HotBot	3.42			

【引用文献】
1) 吉田茂樹．これがインターネットの世界だ．INTERNET Magazine. No.1, p.42-55(1994)
2) 原田昌紀．サーチエンジン徹底活用術．東京．オーム社，1997. 250p.
3) フレッシュアイ―検索が変る 2., http://www.fresheye.com/etc/szuba/pre_0308.html
4) WebSideStory, http://www.websidestory.com/
5) 古関義幸，福島俊一．新世代検索ポータル技術．2001 年情報学シンポジウム講演論文集，http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/sigfi/koseki.pdf
6) Search Engines Take Quantum Leap: 19 out of 20 Now Use Link Popularity To Determine Relevancy, http://www.webseed.com/page1007.html
7) Broder,Andrei etc. Graph Structure in the Web. Proceedings of the 9th World Wide Web Conference. 2000.