WWW 日本語サーチエンジンの比較

池内 淳
atsushi@slis.keio.ac.jp

Ⅰ. はじめに

近年、インデックス・サイズ、検索機能、運営の継続可能性等の諸側面から比較可能な競争力のある日本語サーチエンジンの数が増加している。WWW 上での情報探索を行う際のツールとして、また、代替的情報源として、どういったサーチエンジンを用いるべきであるかについては、これまで、しばしば議論されてきた。その一方で、とくに、日本語ページを主な収集対象としているサーチエンジンに関しては、実際に、サーチエンジンを選択する際に有効と思える情報が充分に存在しているとは言い難い。そこで、本調査では、九つの日本語サーチエンジン（ロボット型全文検索）に対して、辞書・事典類から無作為に抽出した 140 のキーワードを用いた検索を行い、その結果を用いて、多様な観点から、それらサーチエンジンの定量的な比較を行った。

Ⅱ. 既往調査

サーチエンジンの比較・評価を行った研究は、これまでに数多く存在し、その多くが Web 上で公開されていることは周知の通りである。また、このトピックに関する網羅的なリンク・ページも複数存在する [1]。

さて、サーチエンジンの比較・評価のためのアプローチとしては、検索サイト自体を情報源とするか、もしくは、その運営主体によって公表されている情報を利用する場合と、第三者が何らかの手だてを講じて調査を行う場合とに大別されるが、本研究では後者のアプローチを採用する。

比較・評価の対象となる項目についても様々なものが存在する。例えば、

⑴ 検索機能の比較
⑵ 出力件数のカウント
⑶ インデックスサイズの比較・推定
⑷ 重複率（ユニーク率）調査
⑸ サーチエンジン間の一致率
⑹ 適合文献との照合
⑺ 人気度調査

などが挙げられる。このうち、⑴ や ⑵ は極めてポピュラーであり、一般の図書・雑誌だけでなく、Web 上においても、その種のページは頻繁に見受けられるが、サーチエンジン自体の変化に呼応して、つねにデータを更新しているサイトは少ない。

また、⑶ や ⑸ については、強く学術的な関心を喚起しており、これまでに、サーチエンジンのインデックス・サイズ、索引化可能な Web ページ数、及び、それに対するサーチエンジンのカバー率、などを第三者が推定するための手法が幾つか考案されている [2]。

一方、⑹ は情報検索研究の分野では親和的なアプローチであるが、既存の調査については、サンプル数の少なさなどから、その調査結果が他所において言及されることは少ないようである。また、これに準ずるものとして、著名な企業・団体名等を検索した際に、そのオフィシャル・サイトが上位にランクされるか否かといった実験も存在する。

⑺ については、Media Metrics や Nielsen などの調査会社によるもの、アクセス・ログを取る方法、アンケート調査など様々な方法がある。

このほか、言語処理の問題がクローズ・アップされる場合もあり、とくに、日本語環境

では、仮名漢字の異同の認識などに関する調査が存在する。また、言うまでもなく、検索アルゴリズムやページ収集の方針などに関する著述もしばしば見受けられる。

以上のように、サーチエンジンの比較･評価方法は多様であるが、冒頭でも述べたように、日本語サーチエンジンのみを扱った事例は少ない。なかでも、実際の検索を行う際に、各サーチエンジンによって得られる結果にどのような差異が存在するのかに焦点を当てた調査はほとんど行われていないと言える。そこで、本調査では、様々な観点からサーチエンジンを定量的に比較し、それらのデータを元に、各サイトのマッピングを行って、その差異を可視化することを試みる。

Ⅲ. 方法論と結果

本調査では、まず、多様な分野･関心領域を反映する13の辞書･事典類（主に学術的なもの）から、それぞれ10件ずつ無作為にキーワードを抽出し（広辞苑のみ20件を抽出）、全部で 140 のキーワードを用意した。

次に、「検索デスク」[3]をはじめ、既存の日本語サーチエンジンに関する雑誌記事等を参照し、知名度･規模などの観点から、充分に比較可能であると考えられる以下の九つのサーチエンジンを選択した（アルファベット順）。

表 1. 調査対象サーチエンジン

①AltaVista／②Excite Japan／③Fast／④goo／⑤Google（β版）／⑥InfoNavigator／⑦Infoseek Japan／⑧kensaku.org／⑨Lycos Japan

調査は2000年8月16日に行われ、各々のサーチエンジンに対して、140 のキーワードを用いた検索を行った。この際、検索対象としたのは日本語ページ（jp ドメインではない）のみであり、可能であれば、フレーズ検索を行った。このとき、 全ての検索について、上位 100 件までの出力 URL を保存した。

A. サイズの比較

まず、140 件のキーワードのうち、全てのサーチエンジンにおいて 0 ヒットであった 2 件を除外した。次に、調査当時、Infoseek は 40,000 件以上の出力件数を表示しなかったことから、40,000 件以上であったキーワードは除外した。また、九つのサーチエンジンのうち、出力件数が最小であるものが、最大であるものの 5%に満たなかったものについては、インデックス･サイズ以外の日本語処理等の要因に強く負うていることが確かめられたため、併せて除外した。その結果、140 件のうち 89 件のキーワードについて、重複率を勘案した上で、各々の推定出力件数を測定するとともに、各々のキーワードごとに最大出力件数を 100 とした場合の相対値（＝検索力）を算出した。さらに、89 件の平均値を、最も規模の大きかった kensaku.org を 100 とした場合の相対値に置き換え、同時期の「検索デスク」[3]による検索力と比較を行った（→表 2）。

表 2. 相対サイズの比較

	本調査：8/16		検索デスク：8/20	
	検索力	順位	検索力	順位
kensaku	100	1	100	1
Infoseek	95	2	84	2
Lycos	92	3	74	4
Alta	85	4	81	3
Google	84	5	67	5
goo	56	6	54	6
InfoNavi	49	7	39	8
Excite	48	8	38	9
Fast	42	9	47	7

両者の結果はほぼ一致していると言えるが、本調査では、重複率を勘案していることから、若干の相違が存在する（但し、デッド･リンクやミラー･サイトは考慮外である）。

ここで、URL 検索機能を持つ Fast と InfoNavigator を利用して、各々のデータベースに含まれる総ページ数を得るとともに、上で求めた出力件数の相対値を利用して、重複除去後の各サーチエンジンのインデックス・サイズを推定した（→表 3）。

表 3. 絶対サイズの比較（単位：ページ）

	by Fast	by InfoNavi
kensaku	30,951,645	29,933,235
Infoseek J	29,524,654	28,553,196
Lycos J	28,323,715	27,391,772
AltaVista	26,389,569	25,521,266
Google	25,967,068	25,112,667
goo	17,436,916	16,863,185
InfoNavi	15,143,233	14,644,971
Excite J	14,974,545	14,481,834
Fast	12,945,090	12,519,154

※枠線は基準値

B. 一致度の比較

InfoNavigator は、約 150,000 件以上がヒットした場合、URL を出力しないことから、150,000 件以上であったキーワードを除外した。また、ここでは、出力件数の最小値が最大値の 0.5%に満たないものを併せて除外した。その結果得られた 121 件のキーワードについて、出力された URL の上位 100 件までの一致率を測定した。

分析の対象を上位 100 件までとしたのは、実際に、100 件を超えてブラウジングを続けることは極めて稀であると判断したためである。すなわち、ここでの第一の目的は、サーチエンジンの客観的な一致率の測定というよりはむしろ、実際の検索の状況で、各サーチエンジンによる出力結果に、どの程度の一致が見られるのかを定量的に把握するという点にある。

また、一般に、検索者は、出力された URL のうち、多くとも 20〜40 件程度までしか見ないと言われているが、これは定かでない。加えて、ほとんどのサーチエンジンで、一度に表示される URL の最大件数が 100 件でああったことからも、100 件までとした。

次に、121 件のキーワードを、九つのサーチエンジンのうちの最大の出力件数が
①100 件以下のもの（U-100）
②1,000 件以下のもの(U-1000)
③10,000 件以下のもの（U-10000）
④10,000 件を超えるもの（O-10000）
の四つのグループに分割した。

ここで、U-100 のみは、出力結果の全てを直接的に比較することができるため、他の三グループとは分析の性格が異なってくる。

まず、四つのグループごとに、九つのサーチエンジンによって検索された異なり URL 数を求め、そのうち、一つのサーチエンジンでしか出力されなかったもの、及び、九つ全てが出力したものを測定した（→表 4）。

表 4. 出力 URL の一致

	1engine	9engines	異なりURL
U-100	824 46.7%	1 0.057%	1,766 100%
U-1000	7,266 61.4%	4 0.034%	11,825 100%
U-10000	15,751 76.3%	2 0.010%	20,653 100%
O-10000	22,723 84.0%	1 0.0037%	27,063 100%

四つのグループを通じて、九つ全てのサーチエンジンによって等しく出力されているものは極めて少なく、U-100 においても 1%に満たない。また、U-100 では、一つのサーチエンジンによってのみ出力されるものは全体の半分に満たないが、U-10000 や O-10000 では、大半が単一のサーチエンジンでしか出力されないことが分かる。

次に、U-100 を対象として、異なり URL 数のうち、各サーチエンジンがどの程度をカバーしているのかを計測した（→表 5）。

表 5. サーチエンジンのカバー率 (U-100)

kensaku.org	662	37.5%
AltaVista	615	34.8%
Lycos J	600	34.0%
Infoseek J	551	31.2%
Google	393	22.3%
Excite J	362	20.5%
Fast	306	17.3%
InfoNavigator	303	17.2%
goo	238	13.5%
All engines	1,766	100%

その結果、最大の kensaku.org でも、そのカバー率は異なり URL 全体の 37.5%程度でしかないことが分かる。日本語のみを対象としたサーチエンジンについても、もはや、増加し続ける Web ページを単一のサーチエンジンによって網羅することは極めて困難であることを裏付けている。また、これは Lawrence & Giles[4]による米国のサーチエンジンを対象とした調査の結果得られた数値と酷似している。但し、彼らの調査は 1999 年に行われており、その後、この値はさらに低下していると考えられる。

次に、各グループごとに、サーチエンジン間の URL の一致率を測定した。U-100 では、最大である kensaku.org の場合、他のサーチエンジンの 40～50%程度を常に網羅していることが分かった。また、ここでは、15～50%程度の一致率が見受けられたが、U-10000 や O-10000 では、一致率は 10%に満たないものが殆どであった。

C. 相対出力順位の比較

最後に、U-100 を対象として、各サーチエンジン間で一致した URL の相対出現順位を確認した後に、その順位相関係数を算出し、平均値を求めた。これは、データベースの規模や内容といった要素を出来るだけ排除し、日本語処理を含めた広義の検索アルゴリズムという観点から、各サーチエンジンにどの程度の類似性が見受けられるのかを把握するための手だてであるが、語彙の重み付けの際に、一般に、idf が用いられていることからも、個々のデータベースの内容の影響を完全に排除することは出来ない。

この結果、順位相関係数の平均値の最も高かったものは、goo と Excite Japan であり、その値は約 80%にものぼった。また、無相関に近いものや、平均が負の値を示すものも存在したが、概ね、やや強い正の相関が見られた。

さらに、この結果から、数量化IV類によって、サーチエンジンをマッピングしたものが図 1 である。

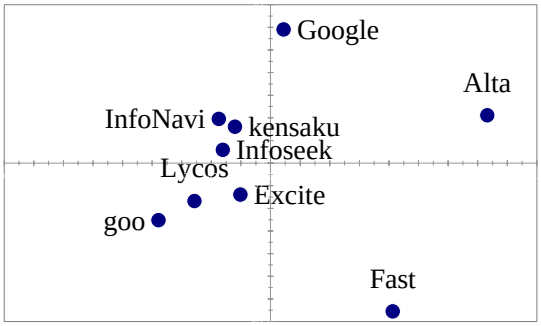

図 1. 相対出力順位によるマッピング

ここで特徴的なのは、日本のサーチエンジンの類似性が高いのに対して、Google、AltaVista、Fast といった米国の多言語サポート検索サイトが、それぞれ独自の位置を占めているという点にあると言える。

＜注･引用文献＞

1) 例えば、Literature about search services. http://www.lub.lu.se/desire/radar/lit-about-search-searvice.html

2) 原田昌紀. WWW ロボットとサーチエンジンのスケーラビリティ. bit. Vol.31, No.12, p.22-28(1999)

3) 検索デスク. http://www.searchdesk.com/

4) Lawrence, S; Giles, C.L. Accessibility of Information on the Web. Nature, Vol.400, p.107-109(1999)