Library and Information Science No. 37 1997

インターネットのサーチエンジンの評価尺度: ESL(Expected Search Length)を使った検索実験

A Measure for Evaluating Search Engines on the World Wide Web: Retrieval Test with ESL (Expected Search Length)

安形 輝, 野末道子, 服部紀彦, 上田修一
Teru Agata, Michiko Nozue, Norihiko Hattori, Shuichi Ueda

Résumé

Search engines are a kind of information retrieval systems on the Internet. However, they cannot be evaluated with the measures commonly applied in traditional retrieval experiments, i.e., recall and precision, since recall cannot be calculated in the essentially unrestricted resources of the Internet. The Expected Search Length (ESL) proposed by Cooper in 1968 is a measure for evaluating information retrieval systems and has the possibility of taking the place of recall and precision. The ESL calculates the cost paid by a user, i.e., number of retrieved sites the user must look through before (s)he gets sufficient number of relevant sites. This article proposes a version of the ESL in order to adapt it to evaluate search engines. An experiment was carried out to compare the validity of the ESL with that of precision and recall. Eight search engines of the domestic and overseas were evaluated by precision and ESL. The result of this retrieval experiment shows the adaptability of ESL for the evaluation of the search engines.

安形 輝: 慶應義塾大学大学院文学研究科図書館・情報学専攻, 東京都港区三田 2-15-45
Teru Agata: Graduate School of Library and Information Science, Keio University, 2-15-45, Mita, Minato-ku, Tokyo. itasan@slis.keio.ac.jp
野末道子: 鉄道総合技術研究所輸送システム開発推進部, 東京都国分寺市光町 2-8-38
Michiko Nozue: 2-8-38, Hikari-cho, Kokubunji-shi, Tokyo. michiko@rtri.or.jp
服部紀彦: 慶應義塾大学文学部図書館・情報学科, 東京都港区三田 2-15-45
Norihiko Hattori: School of Library and Information Science, Keio University, 2-15-45, Mita, Minato-ku, Tokyo.
上田修一: 慶應義塾大学文学部図書館・情報学科教授, 東京都港区三田 2-15-45
Shuichi Ueda: Professor, School of Library and Information Science, Keio University, 2-15-45, Mita, Minato-ku, Tokyo. ueda@slis.keio.ac.jp
受付日: 1998年1月27日 改訂稿受付日: 1998年7月27日 受理日: 1998年9月8日

I. 情報検索システムとしてのサーチエンジンの特色

現在，インターネット上の Web ページの探索には，一般的に「サーチエンジン」と呼ばれるシステムが利用されている。サーチエンジンは，発展途上にあるが，その基本は従来の情報検索システムの延長である。情報検索システムモデルでは，データベースを作成し，データベースを検索用に構造化し，特定の検索手法を用いて，個々の利用者の作成した個別の検索質問に応じて検索し，結果を表示する。

インターネット環境下では，同一の対象に対して異なるサーチエンジンが存在し，競合状態にある。そのため，サーチエンジンが今後発展するためには，これらの異なるサーチエンジンの評価が不可欠であり，既にいくつかの試みがなされている。情報検索システムの一つであるサーチエンジンの評価には，これまでの情報検索システムの評価で用いられてきた手法や尺度を導入することが可能である。

しかしながら，その評価においては，従来の情報検索システムでは考慮されなかった要素を含めなければならない。

第一は，インデックス作成の問題である。これまでの検索実験では，最初に特定のテスト集合を定め，それに対する検索性能を測定してきた。ところがサーチエンジンが扱うのは原則的にインターネット上のデータ全てであり，その意味でテスト集合はインターネット全体と考えられる。多くのサーチエンジンでは，インデックス作成もロボットによって自動化している。従って，これまでは単にテスト集合全体からインデックスを作成するため，その作成手順，収録範囲等は大きな問題ではなかったが，制約の多いインターネット上ではいかに網羅的かつ効率的，さらには高い更新頻度でインデックスを作成できるかが重要である。つまり，サーチエンジンでは，サーバーから Web のページを収集するロボットの機能と性能を含めた評価が必要となる。

第二は，サーチエンジンの利用者がエンドユーザーであるという点である。従来の情報検索システムは，必ずしもエンドユーザーが利用するものではなかった。エンドユーザーのインターネット利用環境は，多くの場合，電話回線でインターネットプロバイダの提供するアクセスポイントに接続する形態をとっている。その場合に，電話回線使用料，インターネット接続料は原則的に接続時間の従量制で決まるため，接続時間を最小としようとする圧力が強く働くと考えられる。サーチエンジンで，検索結果が適切に順位付けがされていれば，検索結果すべてをみる必要がなくなる。結果として，接続時間が短くなるため，検索結果は順位付けがされていることが望ましい。現実に多数のサーチエンジンは，順位付けを行っている。しかし，情報検索における伝統的な評価尺度である精度と再現率は順位付け出力が考慮される以前の尺度であり，適切な評価を行えない恐れがある。

こうしたサーチエンジンの特色を考慮するならば，情報検索システムの伝統的な評価尺度である精度や再現率に代わる別の新しい評価尺度を考案

する必要があろう。以下では，これまでの評価手法と尺度を概観し，それらを検討した後，ESLと呼ばれる評価尺度に改良を加えた尺度を提案し，その有効性を確かめるために，8種のサーチエンジンの評価を試みる。

II. 情報検索システムの評価と評価尺度

A. 従来の評価研究

情報検索システムに対する評価は，伝統的に再現率(recall)と精度(precision)という尺度を使って行われている。これらの尺度は，Aslibによるクランフィールド(Cranfield II)プロジェクトにおける実験に端を発する。この実験では，どの程度のレレバント文献が検索されたか，どの程度非レレバント文献を検索しなかったかという点から評価され，前者は検索の網羅性を測る“再現率”として，後者は検索の正確さを測る“精度”として提案された。

再現率と精度による評価は，一般的にはクランフィールド型検索実験と呼ばれ，具体的には以下のような手順から構成される。

1) テスト集合と検索質問を用意する
2) 各検索質問に対して，どの文献がレレバントであるか，レレバンス判定を行う
3) 情報検索システムはテスト集合に対して検索を行う
4) 検索結果とレレバンス判定の情報を使い，以下の式から再現率と精度を算出し評価を行う

$$\text{再現率} = \frac{\text{検索されたレンバントな文献数}}{\text{全てのレンバントな文献数}}$$

$$\text{精度} = \frac{\text{検索されたレンバントな文献数}}{\text{検索された文献数}}$$

現在までに多くの研究で，クランフィールド型検索実験には多くの問題があることがたびたび指摘されてきた[1)2)]。再現率と精度に焦点を絞り，問題点をまとめると以下のようになる。

1) 再現率算出には全レレバント文献の情報が判明している必要がある。そのため，現実の情報検索では再現率の算出は非常に困難である
2) 再現率と精度が検索されたかどうかを問題にするために，順位付け出力を適切に評価できない
3) 1950年代のバッチ形式の実験環境における尺度であり，インタラクティブな情報検索システムへの応用可能性については疑問が残る
4) 評価には再現率，精度両方が必要であるが，二つの尺度は反比例の関係にあり，評価結果の解釈を困難にしている

サーチエンジンの評価においては，このような問題点のなかでも1) 2)は致命的であり，再現率と精度両方による評価は困難であると言える。実際，後述のように再現率と精度両方からサーチエンジンを評価しているものはない。そこで，従来の情報検索評価研究において再現率と精度に代わるものとして提案はされたが，ほとんど使われてきていない尺度についても検討する。

情報検索評価研究において再現率と精度以外の評価尺度として現在までに様々なものが提案されてきた。それらの評価尺度の中で代表的なものとしては，A. R. Meetham[3)]によるCTI (Contingency Table Information), J. A. Swets[4)]によるArea, J. W. Wilbur[5)]によるレレバンス情報(IdF), H. P. Freiと P. Shauble[6)]によるuと u*, W. S. Cooper[7)]によるESLがある。また，一部のオンライン検索システムに関する研究においては，利用者の満足度(satisfaction)や有用性(utility)といった基準から評価が行われることもある。これらの評価尺度が再現率・精度に置き換わらなかった理由としては，全ての尺度が何らかの問題を抱えており，少なくとも再現率・精度と同程度の妥当性しかなかったことがある[8)]。例えば，Wilburのレレバンス情報は，その算出には再現率と同様に集合中の全レレバント情報が必要であるという問題を抱えている。

全体的な傾向として，再現率と精度を含めほとんどの評価尺度は，検索結果から得られる情報あるいは便益を測定する点で，出力面を扱っているといえる。一方で，検索結果を吟味する場合のような利用者の労力，時間といった利用者側のコス

トは，固定としたり，あるいは無視している。唯一，ESL のみが，逆に利用者の満足する件数の形で得られる情報量を固定し，利用者側のコストを評価する尺度となっている。

B. Cooper による ESL[7)]

Cooper による ESL は，簡単に言えば，利用者の満足する文献数を得られるまで検討しなければならない非レレバント文献数である。そのため，利用者の満足する件数が明らかであることを前提としている。ESL の概念を説明するために図 1 を使う。この図は検索システムにおける順位付け出力の結果を表しており，四角形は文献を，その横の数字は順位を示している。検索質問に対するレレバント文献は灰色で示している。その検索質問に関して利用者が満足する文献数が 3 件だとするならば，その 3 件を利用者が得るまでに利用者が見なくてはならない非レレバント文献は 2 件であるので，ESL は 2 となる。

実際の検索では複数の文献が同順位に出力される場合もあるため，そのことを考慮に入れた ESL は以下のような式 (1) から算出される。

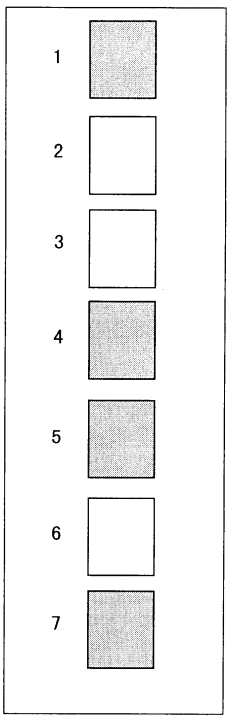

図 1　順位付け出力の例

$$esl(q)=\sum_{m=1}^{M} pr(l_m)\cdot l_m \qquad (1)$$

この式は検索要求 q に対する ESL を表している。この式で，M は同順位の文献が複数あるとき利用者が検討する順序の場合の数である。例えば，満足するレレバント文献数が 4 であり，順位 5 までに得られたレレバント文献数が 3 であり，順位 6 が 3 文献 A, B, C であったとき，その検討する順序には A–B–C, A–C–B, B–A–C, B–C–A, C–A–B, C–B–A の 6 つの状況が考えられるので，この場合 M は 6 である。l_m はある状況 m における検討すべき非レレバント文献数であり，$pr(l_m)$ はその状況になる確率である。ここで，要求が満足される文献数に到達する順位までは単に非レレバント文献の数を考慮すればいいことになるので，実際の計算は以下のような式 (2) で行われる。

$$esl(q)=j+\frac{i\cdot s}{r+1} \qquad (2)$$

ここで，j は該当順位前までに出現した非レレバント文献数であり，i は利用者が満足する件数が得られる順位に存在する非レレバント文献数，r は該当順位に存在するレレバント文献数である。この場合にレレバント文献同士の間に非レレバント文献が何文献あるかを考えるとその期待値は $i/(r+1)$ となる。ここで s は該当順位の何件目のレレバント文献で利用者が満足するかを示している。つまり，$i\cdot s/(r+1)$ は該当順位におけるレレバント文献が満足する件数に達するまでに検討する必要のある非レレバント文献となる。

実際の情報検索システム間の比較を行う場合には，多くの検索式に対する ESL を算出し，その平均を出すために以下の式を使うことになる。

$$\overline{esl}=\frac{1}{N}\sum_{n=1}^{N} esl(q_n)$$

$$=\frac{1}{N}\sum_{n=1}^{N}\left(j_n+\frac{i_n\cdot s_n}{r_n+1}\right) \qquad (3)$$

この式で N は用意した検索質問数を表しており，その他の変数は式 (2) と同様である。

さらに，Cooper は，複数の実験環境において

テスト集合，検索式が異なる場合にも比較を行うことができるようにESLをさらに拡張したESL減少係数を提案している。しかし，この尺度は再現率と同様に集合中の全レレバント文献数を必要とするため，インターネット上のサーチエンジンには応用することができない。実際にサーチエンジンの評価を考えた場合には，それぞれのエンジンについて複数のテスト集合があるわけではなく，インターネット全体は一つの大きな集合と考えられる。従って，ここでは，比較を行うためにはESL減少係数でなく，ESLを使うことが妥当と考える。

ESLの特徴としては，1) 再現率・精度とは異なり単一の尺度である，2) 利用者の満足する件数を考慮に入れている，3) 利用者のコストに焦点を当てた尺度である，ことが挙げられる。

C. ESLの問題点とその改善

ESLには，利用者の満足する件数があらかじめ分かっている必要があるという欠点がある。現実の情報検索行動においては，事実検索ではこの満足する件数は容易に決定されうるものであるが，主題検索のような場合には明確に決まらない場合も存在する。

ESLに関するCooperの研究は多く引用されているが，この欠点のために，ESL自体はほとんど使われてこなかった。ここでは，縦軸にESLを取り，横軸に利用者の満足する件数を取るグラフを考えることで，その欠点を解消することを提案する。例えば，ある仮想的な環境において2つの情報検索システムA, Bに対するESLの平均と利用者の満足する件数が図2のようになる場合を考える。ESLの値は少なくなるほど性能が高いことになるが，この図ではシステムA, BのESLは件数によって高低がある。この図からは，利用者の満足する件数が少ない場合にはシステムAの性能が高く，多い場合にはシステムBの性能が高いことが分かる。

このようなグラフを用いて評価を行うことで，ESLの欠点を解消するとともに，利用者の満足する件数の変化に伴った詳細な分析を行うことができる。

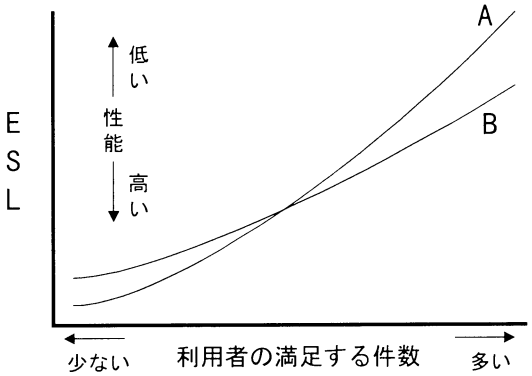

図2 ESLと利用者の満足する件数

本研究と同様にCooperのESLを取り上げている1997年のM. D. Dunlopによる研究[9]でも，ESLのグラフ化手法が提案されている。これはESLに対してシステムごとの検索時間の算出を行い評価することを提案しているが，各システムに応じた複雑な操作を行っているが故に，ESLの利点である汎用性，単純さを失っていると指摘できる。しかし，この研究はコストに焦点を当てたESLという尺度が近年見直されてきたことを示すものと考えられる。

D. ESLと精度の関係

ESLと従来使われてきた精度は，ESLが検索された文献中にある非レレバント文献数，精度が検索された文献中にあるレレバント文献の割合を表すという点で，同じものを表現しているとも考えられる。しかし，以下の点で二つの尺度の概念は異なるものである。

1) ESLは利用者のコストを表すのに対し，精度は正確さを表している
2) ESLは単一の尺度であるのに対して，精度は再現率との関係で用いられる

ESLと精度の違いをより明確に示すために，ある仮想的な情報検索システムA, Bによる極端な検索結果の評価を考える。システムA, B共に検索結果100件中のレレバント文献数が50件であるが，システムAは順位1〜50にレレバント文献が集中し，システムBは順位51〜100にレレ

バント文献が集中しているものとする。精度は100件出力された時点で評価することとし，ESLでは利用者の満足する件数を50件としている。この場合，明らかにシステムAの方が優秀なシステムであるが，精度で見るとシステムA, B共に50%と評価されてしまう。一方，ESLにおいては，システムAは0件，システムBは50件となり，ESLではシステムAの性能が高いと評価される。これはESLが順位付け出力を考慮に入れた尺度であるためである。

III. サーチエンジン評価の問題点

近年，インターネット上の情報資源，なかでもWWW (World Wide Web) による情報は激増しており (http://www.mit.edu/people/mkgray/net/internet-growth-summary.html), 将来的にもさらに増加の一途を辿ると考えられる。情報資源の増加は，有用な情報だけでなく，不要な情報の増加も意味している。そのため利用者がインターネット上の情報の中から必要な情報を選択することは次第に困難になりつつある。そこで，必要な情報を重要と思われる順に提示するサーチエンジンの重要性が高まっている。それとともに，サーチエンジンの評価に関する研究もここ数年数多く行われている。Louise T. Suはインターネット上でしか公開されていないような評価研究も含めレビューを行っている[10]。

従来のサーチエンジンに対する研究では，多くの場合検索性能ではなく，ある検索語によって検索された件数，インターフェースの使いやすさ，検索機能，応答時間，収録数，更新頻度といった点から評価が行われてきた。例えば，国内のサーチエンジン評価研究において著名な浅井は「検索力」という評価尺度を提案している。これはある決められた検索語で何件検索されるかという非常に単純なものである[11]。

しかし，サーチエンジンは情報検索システムであり，情報検索システムは利用者にとって必要な情報を取捨選択することを目的とするものである。従って，当然，何らかの形でその検索性能は評価されなければならない。

従来のクランフィールド型検索実験で検索性能は再現率と精度から測定されてきたが，サーチエンジンの場合には，精度は算出できるが再現率は算出できない。インターネット上の資源は膨大であり，再現率を算出するためにすべての情報を吟味することは不可能なためである。実際に，サーチエンジンの評価を行っている研究では，再現率は放棄されているか省略されている[12) 13)]。その結果として，サーチエンジンに関する評価研究においては精度のみが使われているが，元来再現率と精度は両方の関係から評価を行うものであるため，精度だけによる評価では十分とは言えない。また，サーチエンジンのほとんどは順位付け出力を行うが，前述のように再現率，精度は順位付け出力には対応できない尺度であり，あるシステムが高順位にレレバントなWebページを多く出力したとしても，それが適切に評価に反映できないという問題がある。

つまり，従来のサーチエンジンの評価に関する研究では，検索性能以外の尺度から評価を行っている場合がほとんどであり，さらに，検索性能に関する評価が行われている場合でも，精度しか使われていないため，十分な評価が行われていない。

また，インターネット上の検索では実際に利用者が電話などの低速回線を使って接続していることが多い。このため，利用者は今までの情報検索のように検索の正確さや網羅性というよりは，経済的時間的コストをいかに減らせるかにより重きをおくと考えられる。

本研究ではESLという情報検索領域でほとんど注目されてこなかった尺度を取り上げ，その特徴を生かしながら欠点を改善する評価手法を提案している。この評価手法は，他の評価手法と比べコスト面に焦点をあてている点が特徴であり，よりサーチエンジンの特性にあった評価を行うことができると考えられる。

IV. ESLを用いた評価実験

CooperのESLを使って実際のサーチエンジンを検索し，各サーチエンジンの性能について評

価を行った。その実験の結果からESLの有効性について検討した。

A. 実験環境

実験対象としたサーチエンジンは，国内4種，海外4種の合計8種である（表1）。これらは順位付け出力のできること，1997年秋の時点で代表的なサーチエンジンであること，できるだけ海外版と国内版があることを条件として選択した。さらに，順位付け出力は行えないが，代表的なディレクトリサービスであるYAHOO!, YAHOO! JAPANの二つを参考サービスとして加え，精度を使っての比較評価を行った。実験は1997年10月15日から11月24日の時期に行った。

検索質問は先行研究[14]で使われた質問に基づき国内外それぞれ6件ずつを用意した。ただし，質問の1件は国内のサーチエンジンでの検索結果が少なすぎ除いたため，国内5件，海外6件の合わせて11件を使って評価を行った。サーチエンジン検索に使われた6件の検索質問は以下のようになっている。

- ギリシャ哲学 (Greek philosophy)
- 遺伝的アルゴリズム (genetic algorithm)
- ボランティアの募集 (volunteer)
- 自己破産 (voluntary bankruptcy)
- 長野オリンピック (Nagano Olympic)
- 世界の七不思議 (Seven Wonders of the World) [海外のみ]

各検索質問に対し実際にサーチエンジンを検索する際には，各サーチエンジンで用意されている機能，文法に適した形で検索式を構築した。調査者はインターネットの利用経験が3年以上あり，各サーチエンジンに習熟していることを条件として選択した3名とした。レレバンス判定は調査者が行った。各質問に対する判定者は1名であり，これは判定方針を一貫させるためである。また，現実のレレバンス判定では，情報検索過程の進行につれ，判定基準の変化することが知られている。変化しては一貫した判定を行うことができなくなるため，最初の検索式構築のさいに，その検索質問に対する判定基準を記述しておくことで，判定基準が変化することを防いでいる。

表1 実験対象サーチエンジン

	対象サーチエンジン	参考サービス
日本	goo, Infonavigator, Infoseek Japan, Excite Japan	YAHOO! JAPAN
海外	Alta Vista, HotBot, Infoseek, Excite	YAHOO!

レレバンス判定は，各サーチエンジンの検索結果に対して，リンクを辿った最初のページの情報から行った。1ページ単位で判定することとした理由は，サーチエンジンにおける情報単位のとり方と，時間的な制約の二つである。第一に，現在，サーチエンジンにおける情報の単位は物理的な1ページである。そのため，システムの意図された動作を評価するならば，評価も1ページ単位で行うことになる。第二に，WWW上の情報，およびサーチエンジン内のインデックスは，ある程度の期間で更新されてしまう。そのため，各質問について複数のサーチエンジンの検索結果を比較する際には，できる限り短期間で行うことが望ましい。ここでは，各質問について1サーチエンジンすべての検索結果を1日でレレバンス判定することとした。各質問で検討した件数は多いもので200件以上にも及んだ。回線状況にもよるが，すべての検索結果についてページ中のリンクも辿り，複数ページを見ることは現実的に難しいため，ここでは検索結果の判定を最初の1ページのみで行うこととした。

実際のWebページには，1ページで内容が完結しているページもあれば，他のページと組み合わされてある内容を表現しているページも存在する。今後は1ページという物理的な単位でなく，内容的な単位から扱うサーチエンジンが登場すると思われる。そのようなサーチエンジンを評価するさいには，Webページにおける情報の単位をどのように規定するべきかを検討する必要があるだろう。

判定のレベルは，レレバント，非レレバントの2段階とした。そのページに存在するリンクを辿れば，レレバントな情報が入手できると予想され

検索質問：Genetic Algorithm (遺伝的アルゴリズム)
検索式例：AltaVista -- “genetic algorithm”
判定基準：遺伝的アルゴリズムの概要、定義、使い方、実現方法のいずれかについて書いてあるページをレレバントとする。遺伝的アルゴリズムという語が単に載っているページは除く。また、説明なしに計算・図があるものは除く。
実験結果：

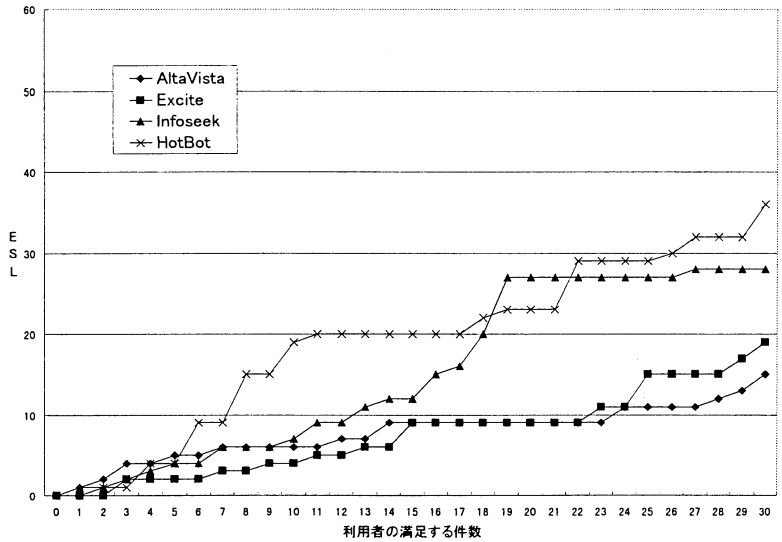

図3　検索質問「遺伝的アルゴリズム」

る場合には，レレバントとした。検索結果の出力に対するレレバンス判定は，各質問，各システムについて利用者の満足する件数が30件に到達し，さらに精度の算出点である検索件数が50件になるまで行った。検索結果に重複があれば，利用者にとってはその分検討する必要のあるページが増えると考え，上位に既に出力されているページはすべて非レレバントとした。各エンジンの検索結果からページまでたどり着けない場合には，インターネット上のトラフィックが混雑しているなどの理由でタイムアウトしてしまう場合と，検索結果中のリンク情報が古くリンク先が存在しない場合とがある。タイムアウトの場合は，しばらく時間をおいて再びアクセスを試みた。リンク先が存在しない場合には非レレバントとした。

以上のような形で実験を行った検索質問「遺伝的アルゴリズム」の例を図3に示す。

B.　実験結果

各検索質問に関する実験結果の平均をグラフで表現したものが，国内のサーチエンジンに関する図4と海外のサーチエンジンに関する図5になる。この図では下にあるほど性能が高いことになる。

図4と図5において明らかなのは，海外のサーチエンジンの方が国内のサーチエンジンと比較して，全体的に性能が高いことである。この理由としては，1) 形態素解析，構文解析等において日本語は処理が難しいこと，2) 国内のWebページ全体の質が海外のWebページと比べて低いこと，が考えられる。前者の問題は，特に形態素解析の面で顕著である。これは英語では単語が空白によって区切られているが，日本語は何らかの手法で単語を分ける必要があるために，起こる問題である。あるサーチエンジンでは，検索式を処理するさいに細かすぎる単位まで区切ったことが原因

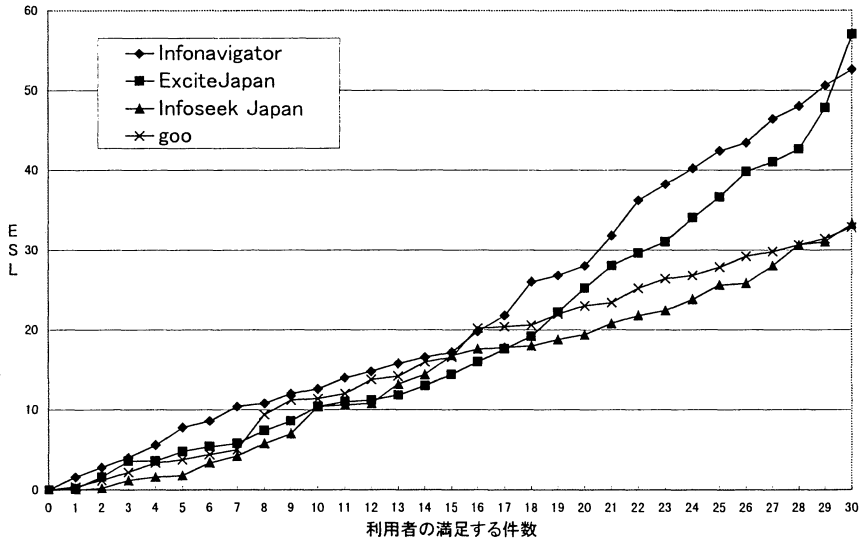

図4 ESL: 国内のサーチエンジン

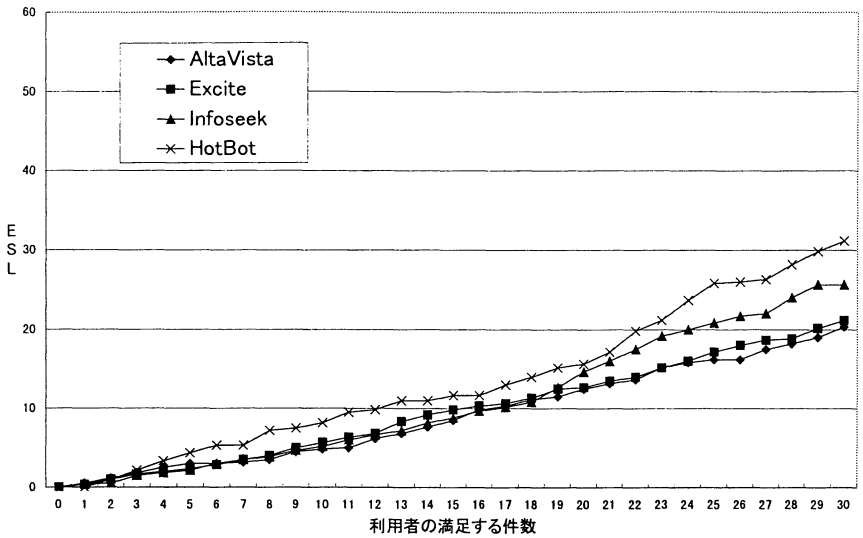

図5 ESL: 海外のサーチエンジン

と考えられるノイズが含まれていた。

今回の実験ではレレバンス判定のデータしか収集してないため，後者の問題はあくまで印象である。しかし，日本語ということで国内の利用しか想定していないためか，情報を公開する意識の低いページが国内には多く存在していた。例えば，ページ内の画像へのリンクが切れているといった単純に技術的な問題のあるページや，「ホームページを作ってみました」という程度の内容の乏しいページなどがあった。

図4から国内のサーチエンジンについて検討すると，必要なレレバント文献数が20件程度までの検索においては各エンジンとも大差ないが，20件以上必要な検索においてはInfoseek Japanあるいはgooを利用した方が検討する件数が少なくてすむことがわかる。次に図5から海外の

表2　50件の時点での精度：国内サーチエンジン

	Infonavigator	Excite Japan	Infoseek Japan	goo	YAHOO! JAPAN
精度（50件）	38.8%	45.2%	50.8%	50.0%	44.2%

表3　50件の時点での精度：海外サーチエンジン

	AltaVista	Excite	Infoseek	HotBot	YAHOO!
精度（50件）	60.0%	59.0%	55.3%	50.7%	55.6%

サーチエンジンについて検討する。どの時点においてもHotBotを使うと他の三つのエンジンと比較して検討するべき件数が多い結果となっている。つまり，海外のサーチエンジンに関しては全体的にHotBotの性能が悪いと考えられる。

各サーチエンジンについて50件の時点での精度を示したのが表2，3である。この表でYAHOO!とYAHOO! JAPANについては，検索件数が全体的に少なかったため，検索結果が50件以上のものはその時点での精度，50件に満たない場合には全検索結果に対する精度となっている。

V. サーチエンジン評価におけるESLの有効性

A. ESLの有効性に対する検討

今回の実験によって，図4のように必要な件数が少ない場合にはどのエンジンを使ってもほぼ同じコストしかかからないが，必要な件数が多い場合にはInfoseek Japanとgooは他の二つに比べると性能が高くなるというように，ESLによってサーチエンジンごとの性能の差異を明らかにできた。つまり，ESLによってサーチエンジンの評価を行うことができると考えられる。

ここでは，ESLの有効性をさらに検討する上で，従来の代表的な評価尺度である精度との比較，さらには再現率を含めての比較も考える。

精度に対して，ESLは必要な件数に応じた評価を行うことができる点が優れていると言える。精度は検索件数を横軸に取ることでESLと同様にグラフを作成することも可能であるが，これには二つの点で問題がある。まず，精度は再現率との関係で考えられた尺度であり，グラフ化する場合には再現率を横軸に取るべきであるという点である。二つ目としては，横軸に検索件数を取り精度をグラフ化したとしても，それが利用者の役に立つのかということがある。つまり，利用者が情報検索システムの評価をシステム選択の参考とする場合に，検索を行う前には，どの程度の分量の情報が欲しいかはある程度予想できるが，何件検索するかは予想できないため，ESLのグラフと比較すると解釈がより困難であるという点である。

また，再現率をも含めてESLとの比較を考えると，評価の基準が，従来の再現率と精度では検索の網羅性と正確さであるのに対して，ESLでは検索にかかるコストである。主に研究のために用いられてきた従来の文献検索では検索の網羅性に対する要求は高かったと考えられる一方で，用途が研究に限られないサーチエンジンひいてはインターネット利用においては，網羅性や正確さというよりはむしろコスト面に対する要求が高いと考えられる。この点でも，サーチエンジン評価においてはESLの方がより妥当な尺度として考えることができる。

さらに，再現率と精度では相反する関係にある二つの尺度をペアにして評価を行うため，結果の解釈が時に難しい場合があるのに対して，ESLは単一の尺度であり，その内容もコストという単純なものであるため，解釈が容易である。

B. まとめ

本研究では，Cooperにより1960年代に提案された尺度であるESLをグラフ化することでサーチエンジン評価に応用し，評価実験を行っ

た。その結果，サーチエンジン評価におけるESLの有効性を明らかにすることができた。

ESLの特徴としては，単一の尺度である，単純であり解釈が容易である，コストに焦点を当てた尺度である，順位付け出力に対応している，ことがあげられる。従来の情報検索システムの代表的な評価尺度は再現率と精度であるが，データベースの大規模化，インタラクティブな情報検索システムの登場によって，それらに代わる尺度の模索が情報検索領域の重要な課題となってきている。今回ESLの有効性を示すことができたことから，ESLの元来の目的であった情報検索システム評価一般に対する応用について，ESLのグラフ化も含め再検討していく必要があると考えられる。

引用文献

1) Ellis, David. "The Dilemma of Measurement in Information Retrieval Research". Journal of the American Society for Information Science. Vol. 47, No. 1, p. 23–36 (1996)
2) Su, Louise T. "Evaluation Measures for Interactive Information Retrieval". Information Processing and Management. Vol. 28, No. 4, p. 503–516 (1992)
3) Meetham, A. R. "Communication Theory and the Evaluation of Information Retrieval Systems". Information Storage and Retrieval. Vol. 5, p. 129–134 (1969)
4) Swets, J. A. "Measuring the Accuracy of Diagnostic Systems". Science. No. 240, p. 1285–1293 (1988)
5) Wilbur, John W. "An Information Measure of Retrieval Performance". Information Systems. Vol. 17, No. 4, p. 283–298 (1992)
6) Frei, H. P.; Schauble, P. "Determining the Effectiveness of Retrieval Algorithms". Information Processing and Management. Vol. 27, No. 2/3, p. 153–164 (1995)
7) Cooper, William S. "Expected Search Length: a Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems". American Documentation. Vol. 19, No. 1, p. 30–41 (1968)
8) Tague-Sutcliffe, J. "The Pragmatics of Information Retrieval Experimentation, Revisited". Information Processing and Management. Vol. 28, p. 467–490 (1992)
9) Dunlop, M. D. "Time, Relevance and Interaction Modeling for Information Retrieval". SIGIR 1997 Proceedings, p. 206–213 (1997)
10) Su, Louise T. "Search Engines on the World Wide Web and Information Retrieval from the Internet: a Review and Evaluation". Online and CDROM Review. Vol. 21, No. 2 (1997)
11) 浅井勇夫. "サーチエンジンからみたWebの世界". 情報の科学と技術, Vol. 47, No. 9, p. 453–458 (1997) [実際に「検索力」を使ったサーチエンジン評価は検索デスクで公開している。http://www.bekkoame.ne.jp/~asaisan/]
12) Leighton, H. V. "Performance of Four World Wide Web Index Services: Infoseek, Lycos, Web Crawler and WWW Worm". http://www.winona.msus.edu/is-f/library-f/webind.htm, 1997/9/4
13) Chu, H. T.; Rosenthal, M. "Search Engines for the World Wide Web: a Comparative Study and Evaluation Methodology". Proceedings of the 59th Annual Meeting of the American Society for Information Science, Vol. 33, Baltimore, 1996–10, ASIS, Baltimore, 1996, 127–135
14) Ding, W.; Marchionini, G. "A Comparative Study of Web Search Service Performance". Proceedings of the 59th Annual Meeting of the American Society for Information Science, Vol. 33, Baltimore, 1996–10, ASIS, Baltimore, 1996, 126–140