特集：インターネットのおとし穴

サーチエンジンからみたWebの世界

浅井勇夫*

インターネットの情報検索に関するホームページ「検索デスク」を運営しており，国内や海外サーチエンジンの巡回検索，オール索引，ニュース，コラム，検索力調査，評価などを試みている。その体験から得られたWebの世界について気がついたことを述べる。主な内容は，新大陸(インターネットの)発見，超スピードの世界，土俵は一つ，Web情報の単位，受動的な登録情報収集，階層型分類からネットワーク型分類へ，ロボット収集の幻想，日本語処理の難しさ，検索機能の発展，検索サービスの選択，などである。

キーワード：Web，サーチエンジン，検索デスク，情報検索，ネットワーク型分類，オール索引，検索サービス，検索力，評価

1．はじめに

インターネットの情報検索サイトとして，96年1月に「検索デスク」を立ち上げて1年半経過しました[1]。巡回検索，オール索引，検索調査，検索ニュース，そしてサーチエンジンの評価などを試みています。検索という地味な分野ですが，現在，1週間に1万人，総数30万の人が訪問しています。1,000以上のサイトが検索デスクをリンクしており(逆リンク)，さらに活字媒体の雑誌や書籍にも多数紹介されています。

95年3月にベッコアメのダイアルアップ接続に加入して以来，毎日，インターネットにアクセスしています。情報発信の期間は1年半と短いのですが，従来の数倍の体験をしました。ユーザー層は広く，初対面の人からの電子メールもあり，オープンでグローバルな世界を実感しています。ここでは，日頃感じているインターネットへの見解，Web情報の質，サーチエンジンの特徴など，その現状と将来の展望について述べていきます。

2．新大陸（インターネット）の発見

インターネットは「新大陸の発見」という身近な歴史的現象と対比して考えると理解しやすいようです。中世のヨーロッパでは天動説が常識でしたが，新大陸の発見や世界一周などで人々は地動説を信じるようになっていきます。そして，ヨーロッパからアメリカ大陸への移住が始まり，長い期間を経て今日のアメリカがつくられます。ヨーロッパの人がどのような心境で新天地へ移住したのか，新しい世界をどのように構築していったかなどは興味深いところです。

昨今のインターネットの動きを見ていますと，インターネットは新大陸の発見に匹敵する程の歴史的な現象とみなせます。情報技術だけに目を奪われがちですが，文化，思想，社会，経済などへの影響を見逃すわけにはいきません。インターネットをどうとらえるかは書く人と読む人との二重のフィルターを通るため，過大評価されたり過小評価されたりします。インターネットの見方は十人十色ですが，どう付き合うかは一人一人の問題です。

ヨーロッパの人がアメリカ大陸へ移住し始めたのは約400年前，13州が独立したのは200年前，世界一の大国になったのは50年前です。アメリカ大陸は400年前からずっと人々に夢と希望を与えてきており，現在でも移住する人が絶えません。インターネットが新大陸であることに気づき，原野を開拓して新しい都市を造ろうと試みているのはアメリカです。しかも大統領が率先して旗を振っています。進出が遅れれば遅れるほど厳しい条件の開拓になることを知っているからです。

インターネット大陸への夢や可能性は膨らむ一方ですが，決して楽園でもユートピアでもありません。それらはこれから作り出すものです。インターネットの終焉を言う人がいますが，私はまだ始まったばかりで，現状は原野や荒野を開拓しているにすぎないとみなしています。ルールも法律もない無法地帯とみなせますし，実際にまだ独立していません。

現実の世界とインターネットの世界との対立は倫理面ぐらいです。もちろん情報メディアへは大きな影響を与えています。今後，セキュリティの問題が解決し

* あさい いさお　大阪府立大学工学部
〒599-8231 大阪府堺市学園町1-1
Tel. 0722-52-1161
（原稿受領　1997.07.16）

て電子マネーが普及し始めますと，資本，マネー，税などの経済的な摩擦は避けれなくなります。インターネットの仮想世界と現存国家の諸制度との経済的な面の対立はインターネット大陸に大きな転換点（独立）をもたらすものと思われます。

3．超スピードの世界

インターネットはパソコン，ネットワーク，ブラウザ，マルチメディアなどの技術に支えられています。その技術発展のスピードは速く，多くの人は進歩についていけず，もうお手上げ状態です。皆がギブアップしようとも技術革新は進んでいます。この激しい競争の原動力はシェア（市場占有率）とデファクトスタンダード（標準仕様）です。

ハードもソフトも3ヶ月位でバージョンアップしています。インターネット関連の企業はシェアをとるために技術競争，価格競争をしています。地球規模のインターネットで成功すれば広大な市場を得，富を獲得できるからです。

このスピード感は，アシモフの科学小説にでてくる，アインシュタインの相対性理論があらわす世界のようです。また，新幹線に乗って見る風景と飛行機に乗って見る風景との差のようです。この景色の変化を理解できれば，現実世界とインターネットとの差が理解できるのではないかと思います。

1年間の出来事がインターネットでは3ヶ月位で出現します。このスピード感というか，3ヶ月単位で物事を考えないと，ついてけなくなります。国の予算や会社の決算は1年単位で決めていますが，やはく3ヶ月単位にしないとインターネットとの波長は合わず，厳しい競争に勝てないのではと危惧します。テンポの速さに慣れるために，人間と情報との接点であるブラウザを少なくとも3ヶ月に1回はバージョンアップしたいものです。

4．土俵は一つ

ハードやソフトの世界だけでなくデータであるWeb情報も2～3ヶ月で更新されています。更新の必要のない情報もありますが，多くの情報は更新がなければ陳腐化します。Web情報を利用する場合にも発信する場合にもこのようなWeb情報の特徴を認識する必要があります。

インターネットは機種に依存せずに情報が利用できます。UNIX，Mac，Windowsのテキストは形式が異なっているため互いに読むことはできません。そこでインターネットでは独自のテキスト形式（HTML文）を作り，ブラウザで表示させています。ブラウザはテキストの変換と表示を行なうソフトです。特筆すべきは，テキストの中にハイパーリンクという他の情報を呼び出すコマンドとマルチメディア関係のタグを埋め込んだことです。

マルチメディア関連の技術の進歩は早く，その技術によって作成されるWeb情報の表現方法や利用方法は急速に進展しています。そのため，作成された情報の陳腐化は早く，1年も経てば使えなくなります。従来の技術は互換性を維持しながら発展したものですが，マルチメディア関連の音声，画像，動画などの技術は互換性を保たずにバージョンアップしています。このことは，デファクトスタンダードを得るための競争が激しいことを示していますが，やむを得ないことです。

ブラウザはハードの統一をもたらしましたが，最近はソフトの統一をしています。以前はファイル転送用のFTPや情報検索用のGopherなどを使う場合には別々のソフトを立ち上げなければなりませんでした。最近ではブラウザがその機能を取り込んでおり，Web情報と同様にあつかうことができます。

Web情報を考える場合，現状と大きな相違点があります。質のよいもの，質の悪いもの，倫理的に問題があるもの，論文，そして宣伝用チラシまでが同じ土俵であるブラウザを通過することです。現状のシステムではいろいろフィルターがかけられていて，異質の情報が紛れ込んだりしません。例えば，図書館で広告やチラシはあつかっていません。

インターネットはすべての情報を一元化し，一つの土俵上に表示するため，利用者は戸惑いを感じるようです。現実でもいろいろな情報があり，それを見るか見ないかは個人に任されています。玉石混淆のWeb情報も現実の世界と同様に個人の責任において利用することが求められています。オープンで何もかも呑み込んでいる世界の力強さを感じます。

5．Web情報の単位

情報を考える場合，あつかう情報の単位は大体決められています。書籍なら装丁した1冊，論文なら1タイトルなどです。ところが，Web情報に関しては何も基準はありません。ホームページのトップなのか，その中にある個々のページなのか決まっていません。それにホームページやページの構成基準もなければ，論文のように標題，要旨，本文，参照文献，出典などの基準もありません。

管理のないまま発展しているWebの情報に枠をはめることは不可能です。情報の単位が不確かなため，情報を収集する場合も，それを探す場合も混乱が生じ

ています。検索サービスのデータベースにどのような種類の情報がどのように入っているか，はっきりしないのが現状です。

Web情報のデータベースはレコードの属性が厳密に定義される従来のデータベースとは異なり，属性という概念すらないといえます。それから，Web情報のデータベース化には抄録型と全文型の二通りの方法があります。抄録型は簡潔なタイトルと1～3行の紹介文から構成されており，さらに分類カテゴリやキーワードが付加される場合もあります。また，全文型はタグ部分を除くテキスト部分だけを対象にするものと，タグ部分とテキスト部分のすべてを対象にするものとがあります。

1HTML文を1ページとした場合，情報の単位は3通りあります。第1はページ単位にする場合です。ロボット系が収集する場合ですが，ゴミのような情報も処理する恐れがあります。第2は複数のページから構成されるホームページを1単位にする場合です。情報内容に応じて一つのホームページができていれば問題ないですが，実際には複数の情報内容を含んでいます。第3はホームページの内容が多様なため複数の情報単位に分ける場合で，これは第1と第2の中間に位置します。しかし，この第3のタイプは，労力を必要とするためか，あまり採用されていません。

いずれにしても，検索提供サービスがどのタイプの情報単位で収集しているのかを見極めて利用することが必要です。

6．受動的な登録情報収集

次々にオープンするホームページの内容を伝える手段として各サーチエンジンはURLの登録を受け付けています。登録する人はURL以外にタイトル，紹介文，キーワードなどの情報を提供しますが，その形式は各サービス会社ごとに異なっています。日本的な清潔感からか，多くのサービス会社は自薦の登録情報だけを処理しています。他人のホームページを紹介するのを遠慮しているからです。

すべての情報が登録される場合には自薦情報だけの収集でよいのですが，情報はすべて登録されていません。したがって，自薦登録だけをあつかうところは情報にムラがあり，重要な情報が含まれないケースが生じます。ユーザーによい情報を提供する立場からは能動的な収集が必要です。

登録情報を受け付けてデータベース化するまでには少し日数がかかりますので，到着分を新着情報として自動的に公表するところが多くみられます。本来，新着情報は新規にオープンしたところが公表する場ですが，ここを宣伝の場とみなして，週1～2回，あるいは月に1～2回，定期的に登録を繰り返すところがあります。これをいつまでも放置していると，「悪貨が良貨を駆使する」ように，よい情報が集まらなくなる恐れがあります。

6月末の1週間に到着した新着情報を調査したところ，Yahoo! Japanは3,000，NTT DIRECTORYが2,100，CSJ Index，InfaNavigator，Nippon SEがそれぞれ1,400～1,500，NetPlazaが1,000，NTTの新着情報が600という結果になりました。古参のNTTの新着情報は多数のサイトに紹介されていますが，他とは明らかに差がついてしまいました。インターネットでの栄光は5分間しかもたないといわれますが，われわれの価値観も絶えずリフレッシュすることが必要です。

7．階層型分類からネットワーク型分類へ

登録された情報はカテゴリに分類されます。全分野の情報を対象にした分類カテゴリを提供するサービス会社は9社あります。そして，各社の分類方法はそれぞれ独自な方法を採っています。

大分類項目にあたるものはトップのページに示されます。これを第1階層の分類とし，そこから次のページに示される中分類項目を第2階層，そして階層が深くなるにしたがって第3階層，第4階層などとします。

第1階層の分類項目数は大体12～20です。第2階層の合計は194～332，第3階層は490～856です。多くの提供サービス会社の分類はこの第3階層までですが，Yahoo! Japanだけは例外で，なんと8階層まであります。それから各階層の項目数を加算した総数は729～1,154ですが，Yahoo! Japanだけは7,696です。

収集データ数をカテゴリ総数で割ると，1カテゴリあたりのデータ数が得られます。今回のデータで求めた結果は30～105とばらつきがありました。この数値が大きいと分類システムが危機に瀕しているとみなせます。登録数が毎週1,500位増加していますので，その増加数に応じてカテゴリ数を増やす必要があります。しかし，ほとんどのところはカテゴリを追加していません。

Webの分類方法は電子メディアやリンク特性などから，

(1) 必要に応じて複数のカテゴリを分類する
(2) 情報量の増大には階層の拡張で対応する
(3) カテゴリ内に他のカテゴリへのリンクを張る

などの特徴をもっています。特に(3)のリンクはネットワーク構造をあらわしており，Yahoo! Japanは@記号で表示しています。

日本のWeb情報の分類は2年くらいした経ってなく，多くの提供サービス会社が階層型分類法を採用しています。情報は増加していますので，階層型分類システムではいずれ崩壊します。はやくネットワーク型分類への移行が望まれます。一方，ユーザーもWeb流の分類方法を理解し，分類されたものを有効に利用するのが望まれます。

なお，検索デスクでは各提供サービス会社のカテゴリを10,635項目収集し，50音別に探索できるようにした「オール索引」を作成しています。日本では唯一のサービスであり，各社のカテゴリ化の相異の解消に役立てばと願っています。

8．ロボット収集の幻想

次に，ロボット収集ですが，これも難しい問題を抱えています。大規模なところはHTML文を1日に数百万ページも収集します。Web情報の陳腐化が激しいため，訪問頻度が月から週，週から日へと変化しています。一方，情報を蓄積しているWebサーバーはロボットの訪問で負荷が高くなっています。多くの人が利用することを目的に情報を公開しているのですが，ロボットの訪問は痛し痒しです。

それから，個人が利用できるエイジェントソフトが普及してきました。これを使用すると，特定のURLの情報を自動的に収集し，ディスクに蓄積できます。時間が節約でき実用的ですが，やはりサーバー側に負荷をかけます。不必要な情報を収集しても，情報過多になり消化不良になるだけです。

ロボットが情報を収集するアルゴリズムはいろいろあります。各検索サービスはシステムの能力に応じて，独自のロボット・アルゴリズムを作成し，広く浅く収集しています。すべてのページを収集していると思うのは幻想にすぎません。

例えば，検索デスクは全体で約185ページあります。日本のロボット系検索サービスは11社ありますが，検索デスクを1ページも収集していないところが4社あり，数ページ収集しているところが6社，Infoseek Japanだけは全ページ(？)収集しています。これは検索に必要な網羅性が満たされていないことを示しています。

97年3月にNTTのgooがサービスを開始しました。その検索力は圧倒的で，日本の検索サービスも第2世代に入ったとみなせます。そして，gooに追いつけとばかりにデータ数を一挙に10倍位増やしたところが数箇所あり，激しい競争が始まっています。現在のgooのデータ数は500万件で，多くの検索サービスは30～50万件です。サーチエンジンを効率的に利用する人には，変化するサーチエンジン事情を把握することは必要です。

9．日本語処理の難しさ

従来の検索システムは，キーワード入力など均質で少量のデータをあつかっていました。しかし，Web情報は玉石混淆の全文データを大量にあつかいます。人力に頼ることは不可能でソフトに頼るわけですが，従来使っていた日本語処理ソフトはほとんど利用できないのが現状です。

日本語の検索サービスをする23社について日本語の処理能力を調べました。漢字の「旅」と「旅行」を区別できるのが11，できないのが12です。旅で検索すると旅行や旅館なども検索されてしまいます。京都と東京都も同じです。

次に，カタカナですが，「サーバ」と「サーバー」を区別できるのが13，できないのが10です。「ロバ」を探すと「プロバイダー」が検索されるなどです。頻繁に検索調査をしますが，日本語処理が不完全なシステムの方が検索数が多くなり，したがってよい評価を得るので困ってしまいます。

英字の検索ですが，大文字と小文字を区別しない方が便利です。それから全角と半角が同じにあつかわれるのもよいシステムです。この両者をパスするのが14社あり，パスしないのが9社あります。海外のシステムでは考えられない日本独自の処理です。

最後に，数字処理ですが，半角と全角を同じにあつかうのが12社，あつかわないのが11社あり，後者のうち，ロボット系の3社は2桁の数字は検索しません。これはシステムに負荷がかかるため2文字以下の単語は処理しないためです。

このように日本語処理は難しく，上記の4種類のチェックすべてにパスするのは，23システムのうち6システムしかありません。登録系の検索サービスは分類が主体であり，検索システムはあとから追加された機能です。

10．検索機能の発展

本格的な検索を可能にする検索式が利用できるようになったのはこの1～2年のことです。米国では95年末にAltaVista，96年5月にHotBot，96年8月にUltraseekが相次いでオープンしました。これらは，いわゆる第2世代のサーチエンジンで，データの収集，タグを含む全文データベース化，そして検索式の利用など，従来のサーチエンジンに比べて飛躍的に進歩しています。

日本では97年3月NTTアドとInktomeが提携し

gooを，また97年5月にInfoseek Japanが日本語版Ultraseekをオープンさせました。上記のサービス以外の検索はAND検索やOR検索しか使えません。たくさんの検索サービスがありますが，検索式が使えるか使えないかがポイントです。

検索する上で問題になるのは，検索式を使えない第1世代のサーチエンジンを利用する場合です。各検索サービスごとに検索ルールは微妙に異なっています。以下，気のついたことを記します。

キーワードの入力で問題になるのが，前節の漢字処理のように単語の区切りがあいまいなことです。英単語の場合，空白で区切りがつくのですが，検索ソフトが対応してないようです。例えば，marketと入力すると，market，markets，marketingなど，いわゆるmarket*と同じ処理をする場合があります。

英単語の場合，小文字も大文字も同じに扱いますが，新しい第2世代サーチエンジンは小文字の入力は大文字も小文字も検索しますが，大文字で入力するとそれに一致するものしか検索しません。

2つの単語をスペースで区切って入力するとデフォルトがAND検索やOR検索があり，迷ってしまいます。それに第1世代のサーチエンジンはAND検索やOR検索ができるのですが，実際に検索してみるとAND検索やOR検索を正しく処理していないシステムがあります。AND演算をmin演算，OR演算をmax演算しているようです。あまり問題にされていませんが，サービス側もユーザー側も検索に不慣れなのが原因のようです。

海外のサービスでフレーズが使えるようになったのは第2世代からです。それまでは，例えば，electronic commerceを検索すると，AND検索された膨大な検索結果が得られます。複合語が多いので，その点は便利になりました。

日本語処理にはフレーズ概念がなく，従来の検索システムでは自動的に単語に区切り，AND検索していますが，検索にフレーズの概念を導入することが望まれます。

11. 検索サービスの選択

インターネット上の検索とユーザーとの関係を見ますと，

1) 検索の概念を知らない人
2) インターネットで検索できることを知らない人
3) 検索でYahoo!しか使ったことがない人
4) 複数の検索サービスでAND/OR検索する人
5) 複数の検索サービスで検索式を使う人

などに分かれます。インターネットは自動車のような運転免許証は必要なく，誰でも利用できます。検索も同様であり，何も知らなくても利用できます。4)や5)レベルの検索をしている人はまだ少数派です。

日本の検索サービスを利用する検索や探索の総数は1日に1,000万件以上といわれています。インターネット上には膨大な情報があり，情報は検索サービスを利用して探すのが当たり前になりつつあります。しかも，無料で利用できるとなれば使わなければ損と言えます。

数多くのサービスがありますが，どの検索サービスを利用すればよいか，ということが問題になります。現状では網羅性の観点から，できるだけ複数のサービスを利用するのがベターです。その理由は海外の検索サービスの場合，データの収集数が半年以上も増加しない所があるからです。

検索システムの目安としてデータ数が問題にされます。データ数が多ければ情報がたくさん入っていると考えるのですが，データ数よりもデータベース化の方が重要です。検索デスクではデータ数に代わって「検索力」を提案し，毎週，統計データを求めています。15種類のキーワードで検索した数の合計を求め，日本の場合はgoo，海外の場合はAltaVista，を1,000とした場合の相対的な指数を求めています。

海外の検索サービスの客観的な評価の基準を示します。次の3種類の要素から構成します。

(1) 検索力
(2) 検索機能
  a) フレーズがあつかえる
  b) AND/ORを正しく処理する
  c) 逆引用を求めることができる
  d) マルチメディア技術を検索できる
(3) 検索結果
  a) データベースの更新頻度
  b) 出力画面全体のデザイン
  c) 検索した各ページの紹介方法
  d) ランキング出力

今後，日本の検索サービスに適用する予定です。

最後の項目にあるランキング出力ですが，ユーザーはそこで何が行われているのか全くブラックボックスです。ランキングの妥当性を検討する必要があります。しかし，激しい競争で絶えず改良が行われていますので，調査もすぐ陳腐化してしまいます。

12. おわりに

優れたものも不完全なものも，すぐに過去のものにしていくのがインターネットの世界です。優れたものも新しいものが出現して不完全なものになり，不完全

なものも新しいものに変身して優れたものになります。価値基準をどこにおけばよいのか迷います。

今後，情報がコントロールされたりして，フィルターを通した意図的な情報しか流されなくなる恐れもあります。検索結果のランキング出力が広告費をだしたところが上位を占めたら興ざめです。規制の少ないオープン状況が維持できるよう皆で協力しあわなければなりません。

Web 情報の信頼性は無に等しいものです。日付が入っている文章でも，あとからいつでも修正が可能ですし，引っ込めることもできます。Web では単純なミスなどは責任を問われません。ソフトのセキュリティホールが問題になっていますが，誰も責任を取りません。製造物の場合は PL 法など厳しい世界です。

インターネットにおける検索サービスの経営基盤は弱く，今後の展開は予測不可能です。商売の場合は，物が売れなくて上手く行かなくなるのですが，Web の場合は情報量が増えてシステムが処理できなくなって上手く行かなくなります。すでに先発組の中には撤退するところもでてきています。

Web 情報の増加に応じてデータ収集が行われていれば問題ないのですが，実際にはデータ収集の伸びは止まっています。これは Web 情報を捕捉できない状況が起きていることを意味します。それを打開するには，未来の検索システムに対する投資が必要です。

インターネットの世界は，作っては壊し，壊しては作る開拓時代であることがお分かり頂けたでしょうか。インターネットの落とし穴は Web の世界を現存の世界に，現存する知識や価値観から判断したり陳腐化した Web 情報から判断することです。テンポの速い Web の現状を正しく認識するために知識をいかにリフレッシュさせるかは永遠の課題です。

参 照 文 献

1）検索デスク http:www.bekkoame.or.jp/~asaisan/

Special feature : Downside of Internet. Inside of Web from Search Engine, Isao ASAI (1-1, Gakuen-cho, Sakai-shi, Osaka 599-8231))

情報の科学と技術　Vol.47 No.9（1997.9）抜刷

特集：インターネットのおとし穴

サーチエンジンからみたWebの世界

浅　井　勇　夫