# Web情報の検索ツールSearch Engineについて

○浅井　勇夫*

## Search Engines as Tools for Searching Web Information

ASAI Isao*

Web情報は爆発的に増加している。それに伴い必要なWeb情報を探すことが難しくなっている。そのため，多数の検索ツールsearch engineが開発されている。ここではWeb情報検索に関する最新の動向と問題点を考察する。Web情報の特徴，Web情報の収集，データベース化，検索手法，検索サービス，そして今後の課題などを扱う。〔著者抄録〕

As Web information very sharply increases in amount, the time required to search the Web for information has become much longer. These difficulties have created a great demand for such tools as search engines. In this study, some newest tips to keep in mind and problems that need to be resolved are carefully considered. This paper also gives a brief description of the characteristics of Web information resources, how to search, collect and classify a great amount of those resources, how to perform the building and searching of a database of the resources, and of how to deal with some related problems in the future. 〔Author Abs.〕

＊　大阪府立大学工学部経営工学科（〒593 堺市学園町1-1） Tel. 0722(52)1161
＊　University of Osaka Prefecture, Dept. of Industrial Engineering (1-1, Gakuen-cho, Sakai, 593)

## 1. はじめに

インターネットにおけるWWW（World Wide Web）情報は爆発的に増加している。それに伴い必要な情報を検索することが難しくなり，多数の検索ツールが開発され，提供されている。検索ということでアクセス数が多く，年間のベストWeb賞を獲得したものもある。

Web情報の検索ツールとして脚光を浴びているのはsearch engineといわれるものである。Web情報を蓄積し，索引を付与して，検索に応じるのであるが，これら一連の処理はソフトを使って自動的に行い，無料で提供される。ここでは，Web情報の特徴，Web情報の収集，データベース化，検索手法，検索サービスなど，Web情報検索に関する最新の動向と問題点を考察する。

## 2. Web情報の特徴

インターネットのWWWを利用すると，非常にさまざまなWeb情報を知ることができる。Web情報をデータベース化しようとする人や，Web情報を検索しようとする人は，新しいWeb情報の特徴を理解しなければ，その作成や利用に失敗する。まず，Web情報の特徴について述べる。

### 2.1 Web情報の急増

管理する者がいないことが，インターネットの特徴かも知れないが，Web情報がどれくらい存在するのか誰にも分からない。しかし，大規模に情報を収集しているWebCrawler[1]のデータを使って推定する。

まず，Web情報を蓄積するサーバー数であるが，1994年12月末の1万4,000台が1995年9月末には5万4,000台になっている。4～5か月で倍増，年間で5～6倍である。この増加傾向が続けば，1997年中頃に100万台，1998年末に1,000万台になることが予想される。

次に，Web頁数であるが，サーバー当たり約160頁とみなすと，1995年9月末で約860万頁である。乗数160は時の経過とともに増加するとみなせば，1997年中頃には2億頁，1998年末には25億頁ぐらいになることが予想される。

一方，ユーザー数の増加も同様な傾向を示すとみなせる。これらの予測が1～2年遅くなったとしても，今後3～5年で情報の世界が劇的に変化していくことが理解できる。

### 2.2 新陳代謝が速い

従来のデータベースは，新しい情報を次々に追加していくことにより価値が増大していった。しかし，Web情報は陳腐化が激しいため，新しい情報を追加していくだけではゴミの山を築くことになる。

Open Text[2]は情報収集時のデータを公表している。その割合を求めると次のようになる。

| | | | | |
|---|---|---|---|---|
| (1) | 情報の変更 | 11,768 | 15.1% | 40.7% |
| (2) | 情報の削除 | 10,658 | 13.7% | 36.9% |
| (3) | 情報が同じ | 6,490 | 8.3% | 22.4% |
| (4) | 新しい情報 | 48,993 | 62.9% | 169.4% |
| | 合　計 | 77,909 | 100.0% | 269.4% |

これはリンクをベースに新しい情報もすでに蓄積してある情報も一緒に収集したときの統計データである。これから，所有しているWeb情報の7～8割が数か月で不要になることが分かる。

### 2.3 ルールがない

#### 2.3.1 テキストだけを処理する

カラフルで画像付きのWeb情報は，主にHTMLファイルとGIFファイルから構成されている。データベース化の対象になるのはテキスト系のHTML文書であり，時間とコストをかけて作成したGIFファイルは対象外である。

#### 2.3.2 書誌項目の記述がない

書誌事項に当たる，著者名，標題，出典，発行年月は考慮されていない。HTMLに〈base...〉はあるが，書いてあるのは希である。また，〈title〉はあるが，丁寧に書かれてない。文章の最初の方にある〈heading〉の方が標題を示す場合が多い。

#### 2.3.3 内容はピンからキリまである

Web情報は会議論文のような最先端の情報から，新聞と一緒に配達されるチラシのようなものまで，情報の質の幅が広い。すべての情報がブラウザを通して一元化する新しい世界の出現なので，やむを得ないことかも知れない。

#### 2.3.4 容量も大小さまざまである

本を章ごとに分割すれば容量は小さくなる。容

量が数百～数十KBまで，その幅は広い。

#### 2.3.5 内容は起承転結になっていない

ハイパーリンクを使って簡単にリンクを張ることができるため，個々のWeb情報の内容はバラバラである。今後，Netscape Navigator 2.0に採用された〈frame〉が普及すれば，情報の把握はますます困難になる。

#### 2.3.6 Web情報の構成は自由である

Web情報の構成には原則がなく，自由に行われる。そのため配置替えが頻繁に行われ，同じ情報がいろいろなディレクトリに出現する。

## 3. Web情報の収集

雑誌の購入から始まるデータベースの作成スタイルはWeb情報の世界には見られない。誰もが情報を受発信できる環境のもとでは，全く新しいタイプの収集法が発達している。

### 3.1 自薦・他薦情報の収集

ホームページがWWWサーバーにセットされ，世界中の人に見てもらう機会を持っても，人々に情報の存在を知らせなければアクセスされない。そのため情報発信者は情報を公開したことを積極的に人に知らせようとする。

新着情報やディレクトリ情報をサービスしている所はWeb頁の中に"Add"や"Submit"などのリンクを張り，一般からの自薦・他薦情報を求めている。データベースの作成者はユーザーに有用な情報を提供すればするほど情報が集まってくる。この仕組みはWeb情報の収集手段として重要である。

### 3.2 ロボット・ソフトで自動収集

検索サービスを始めるのに，自薦・他薦情報の到着を待っているだけでは，なかなか情報は集まらず，後発組にとっては辛いところである。そこで考え出されたのは，ソフトを使い，WWWサーバーにある情報をSpiderというロボット・ソフトを使って自動収集することである[3]。

1日に5万頁収集している所もあり，処理マシンを増やせば処理量を拡大できる。2.2項のWeb情報の陳腐化を考慮しなければならず，サーバーを巡回する方法も考慮しなければならない。所有しているWeb情報を廃棄し，短期間の周期で新たに収集するのが最良の方法かも知れない。

サーバー側にとっては，苦労して作成したWeb情報をSpiderが侵入してきてごっそり持っていくのには抵抗があり，排除する動きもみられるが，検索サービスの重要性が認識されるにつれて，暗黙の了解が成立している。

## 4. データベース化

情報の収集は誰にも門戸が開かれており，検索の優劣を決めるのはデータベース化の所である。

search engineも従来からある単純なものから，情報検索の研究論文に見られる斬新な発想の第2世代のsearch engineが登場してきた。

### 4.1 フルテキスト化

新しい情報が入ると標題やコメントを付け，カテゴリーに割り振り，蓄積するというのが初期のパターンである。情報量が多くなったために検索を加えたが，標題やコメントからの情報しか得られない。

HTMLで書かれているWeb情報はそれ自体フルテキストであり，そこから情報を得た方が精度は高くなる。そのために，ほとんどのsearch engineは独自の方法でフルテキストを加工し，Web情報の特徴を抽出する。

### 4.2 ソフトを使用した自動処理

Web情報から特徴を示す索引を得るのに人が介在すればコストがかかる。また，情報の量，質，陳腐化などを考えれば，直接人間がタッチする問題ではない。処理はソフトにまかせ，人間はアイデアを提供するという立場に立つ。

### 4.3 リンク情報の利用

学術論文は参照文献(references)を持つが，それを小規模化したWeb頁はハイパーリンクを持っている。このリンクはHTML内で簡単に識別することができ，利用しやすいばかりか，非常に貴重な情報とみなせる。

文献の場合は，参照文献のreferencesを転置して被参照文献のcitationsを得た。筆者は，1984年の当研究集会において，referationsという新しい概念(参照文献＋文献自体＋被参照文献)を提案

し[4]，その関係を使った知的情報検索について発表してきた。

WWWにおけるハイパーリンクは文献における参照文献と同じ概念である。情報の単位が文献からWeb頁に変更したとみなせば，リンク情報を利用した情報検索は可能になる。新しいタイプの検索サービスの中に，リンク情報を含めた索引づくりをしているものがあり，従来とは異なった検索が可能となっている。

### 4.4 知識ベースの構築

検索を補佐する道具だては，ますます多彩になってきた。人間の知識をデータとして入力して利用したり，あるいは学習機能を持たせて改善するというアイデアが大胆に取り入れられている。特に，2次，3次のリンク情報を利用すれば，情報量は多くなり，知識ベースの構築がしやすくなる。

検索手法とからむが，収集したソースだけでなく，検索の際に発生するデータをいかに再利用するかは今後の課題である。

## 5. 検索手法

検索する際の入力オプションや検索結果に対する出力オプションが多くなってきた。これらの入出力オプションや検索結果は前節におけるデータベース化の処理と関連している。

### 5.1 〈FORM〉を使った検索条件の設定

検索時のオプションは，ユーザー側ではHTMLの〈FORM〉を使って実現する。送られてきた検索フォームに必要事項を記入し，検索ボタンを押すことにより，検索条件を送り返す。

それをサーバー側はperl言語対応のCGI(Common Gateway Interfaces) スクリプトを通して，データベースにアクセスし，その結果を送り返す。また，従来のデータベースはCGIというフィルターを通してWWWから利用可能になる。

### 5.2 検索式の表現力の強化

検索におけるブール演算子OR, AND, NOTを使った検索式の作成方法は検索サービスごとに微妙に異なっている。それにもまして，単語，あるいは単語間の表現力が強化され，明確に定義できるようになってきた。オプションを選択する形式のものから記号を用いるものまでさまざまである。検索式の入力時に迷うことが多いため，きめ細かなオプションの設定が必要である。

### 5.3 検索手法の高度化

検索を支援するための方法が登場してきた。オンライン検索ではすでに取り入れられているものもあるが，今後さらにいろいろな方法が開発されることが予想される。

検索キーの選択時に，データベースの中に入っている情報をあらかじめ表示するものとして，関連した語，類似の語とその頻度，そして類似のスペルを持つ語などがある。また，検索する語やフレーズに重みを付けたりすることができる。

さらに，検索キーに関連した語を付加して検索することもできる。この場合，何が付加されているのか分からないが，知識ベースの優劣が検索結果に影響する。

### 5.4 関連性の高い順の出力

search engineを用いた検索で特徴的なことは，検索結果が関連性の高い順に出力されることである。これは検索に関して大きな意味を持つ。検索により数千のプレインな結果を得た場合，それを利用することは不可能であり，検索結果を絞り込むためにさらに検索しなければならない。

検索結果が関連性の高い順に出力されると，結果が数千あったとしても，上位の数十を探すだけで必要なものが見つかる可能性が高く，検索の絞り込みのようなことは不必要になる。

関連性の定義として，最大の出現頻度を100としたり，あるいはリンク関係を使ったインデクシングを基に関連性を求めるなどsearch engineごとに異なる。複数の検索サービスの検索結果を比較検討して，個々のケースごとに利用できるかどうかを判断しなければならない。検索手法が成長時でもあり，優劣の結論は急ぐべきでない。

### 5.5 Web頁と類似性の高いWeb頁群の出力

最近の新しいsearch engineは，検索したWeb頁に類似のWeb頁群を検索するものがある。筆者の研究したreferations関係を用いる知的検索でも類似性の高い文献群の出力を扱い，日本はもちろんのこと8年前に米国でも発表した。それが

現在のシステムに影響を与えていれば幸いである。

マウスをクリックするだけで類似性の高いWeb頁が得られるようになったが，その内部の仕組みは公表されていない。類似性を求めるにはWeb頁間のリンク関係を利用しなければならないが，リンクの扱い方は4種類あり，そのどれが利用されているかは不明である。日本でもReferations関係を用いたsearch engineの開発が望まれる。

## 6. 検索サービス

大学や企業がボランティアで行っていた検索サービスは，1995年4～6月にベンチャー企業として独立したり買収されるなどして，新たな段階に入った。そして，1995年10月以降に既存のサービスを改善したり，新しい検索サービスが生まれ，検索サービスの第2ラウンドが始まった。

### 6.1 Web情報検索の発展過程

自然発生的に開始されたWeb情報検索は急速に発展している分野である。現在までの情報検索の発展は次の3段階にまとめることができる。

#### 6.1.1 第1段階：search engineのないもの

Web情報に標題，コメント，それから分類をつけたディレクトリ系のもので，簡単な検索を行うことができる。代表的なものにYahooがあり，検索サービスのトップである。自薦・他薦の情報を収集し，階層的なカテゴリーをブラウズすることにより，必要な情報にたどり着く。1995年9月に第2世代のsearch engineのOpen Textを採用するとの発表があり，トップではあるが，大きな変身を遂げようとしている。日本のディレクトリはYahooを見本にしたのか，残念なことにまだ第1段階にある。

#### 6.1.2 第2段階：第1世代のsearch engine

情報収集にspiderを採用し，コメントではなくフルテキストを使って索引を作成する。そして検索結果を関連性の高い順に表示する。このタイプの代表的なものにWebCrawlerやLycosがある。いずれも検索サービスでは2～3位に位置し，大学で始まり，企業として再出発し，最近，検索画面も新画面に移行した。

#### 6.1.3 第3段階：第2世代のsearch engine

第1世代のsearch engineがさらに強化され，新しい機能が加わった。第1世代との相違はリンク情報を利用するかしないかである。リンク情報を含めた索引を作成し，その結果，類似性の高いWeb頁群を検索することが可能になった。このタイプのものは，まだ知名度が低いが，Yahooを採用したことにより，徐々に移行が始まるものと思われる。

### 6.2 無料か有料か

検索サービスは大学や企業のボランティアとして出発したために無料であった。それが相次いで企業として独立したため，検索サービスを維持するのに必要な設備や人件費などのコストをいかに回収するかが問題である。そのため検索の有料化を打ち出す所もあったが，現在では無料にせざるを得なくなっている。

検索はアクセス数が多いため，検索結果に広告を導入すれば広告料が入り，検索サービスを無料に維持することができる。これは，一時的な現象かも知れないが，ユーザーには素晴らしい世界であることに間違いない。

### 6.3 競争の激化

第1ラウンドの御三家Yahoo，WebCrawler，Lycosは，1995年10月時点で，それぞれ週1,000万回以上のアクセスがある。1回当たり2円の広告料でも週2,000万円の収入になる。そのため，検索結果を良くしたり，検索の応答時間を短くしたり，有用な情報を付加したりしてアクセス数に対する激しい競争が始まったところである。

この第2ラウンドの勝者が誰になるか分からないが，情報検索が数年で大きく進歩することだけは確かであり，想像できないような新しい情報検索の世界が訪れることが予想される。

## 7. おわりに

Web情報検索を取り巻く動向を見てきたが，Web情報に関する研究はまだ始まったところである。第2節で見てきたWeb情報の特徴から，Webデータベースは従来のデータベースの延長線上にない。特に，追加することから，定期的な

更新への切り替えは大きな変化である。

また，Web情報はURLで識別しているが，Web情報は記憶場所を容易に移動できるため，同じ情報が重複して入っている場合が見られる。これを回避するためには，Web情報を同定するため識別コードを付加することが必要になる。筆者は文献の識別に関してAPTSコードを提案したが[5]，著者も発行年もない状態では，新しい方法を考えなければならない。

最後に，Web情報の急増している現状で，情報関係者がなすべき緊急の課題を列挙すると，次のとおりである。

(1) Web情報検索システムの開発
(2) 既存データベースをWWWで利用
(3) 情報関係者の教育

来年の予算でとか，予定にないということは許されず，早急に特別な処置をとるべきである。

情報関係者はWWWの世界に一番身近にいながら，そのことを自覚し，また緊急事態であることを認識している人は少ない。WWWの利用技術を1日も早くマスターし，情報部門という枠から飛び出して活躍することが望まれる。

## 参照文献


1) http://webcrawler.com/WebCrawler/Facts/Size.html
2) http://www.opentext.com:8080/omw-stats.html
3) http://web.nexor.co.uk/mak/doc/robots/robots.html
4) 浅井勇夫．パソコンによる引用文献データベースの開発．第21回情報科学技術研究集会発表論文集．日本科学技術情報センター．東京，1984，p.21-31.
5) 浅井勇夫．大規模なリファレーションDBの開発：文献同定のためのAPTSコードの導入．第29回情報科学技術研究集会発表論文集．日本科学技術情報センター．東京，1992，p.273-278.

# 第32回
# 情報科学技術研究集会
# 発表論文集
## INFORUM'95

JICST 日本科学技術情報センター
The Japan Information Center of Science and Technology