日本のＷｅｂ検索・索引サービスの現状と課題

浅井勇夫（大阪府立大学工学部）

Keywords: Internet, Web Documents, search services, index services

１．はじめに

WWW(World Wide Web)の登場で、インターネットは地球規模で発展・拡大しており、日本も同様であることは誰もが周知の事実である。オンラインデータベースを利用している人の中には、情報の質的な面からＷｅｂ情報に疑問を呈する人もいたが、時が経つにつれてというか驚異的なＷｅｂ情報の増加で、この大きな情報革命の波に巻き込まれつつあるのが現状である。

昨年の当シンポジウムで、「人と情報との接点にWWWブラウザがあり、すべての情報はＨＴＭＬを使用したマルチ文書に統一される」という見解を示した[1]。インターネットは外部情報のＷｅｂ文書化を促進しているが、今度はイントラネットという形で内部情報のＷｅｂ文書化が促進されようとしている。現在はいろいろなメディアがＷｅｂ文書という新しいメディアに統一されていく時代であり、情報に携わる人も積極的に参加し、リーダーシップをとることが求められている。

日本語Ｗｅｂ文書は急増しており、その所在を検索・索引サービスする所が増加している。この日本語のサービスは１年前には皆無であり、この半年位の間に誕生したものである。筆者は今年の１月中旬から、プライベートなホームページ「検索デスク」をインターネット上に開設し、検索・索引サービスに関する話題を提供している[2]。ここでは、まず情報発信の一事例として検索デスクを紹介し、日本語のＷｅｂ検索・索引サービスの現状とその課題について述べる。

２．情報発信の一事例

検索デスクは１月 18 日の NTT 新着情報への登録でオープンした。その直後に電子新聞 INTERNET Watch の検索特集に紹介[3]されたのを手始めに、多くのホームページの検索コーナーで紹介され、５月 10 日時点での逆リンク数は 97 件に達している。ホームページの内容は最初のうちは工事中の所が多かったが、何回も改装を重ね、徐々に充実したものになっている。５月の連休中に、フレームと JavaScript を使った新画面に改装した。それと雑誌の紹介が重なり訪問者は増加している。以下、現時点での内容を紹介する。

ASAI, Isao
College of Eng., Univ. of Osaka Prefecture
〒593 堺市学園町 1-1　Tel.0722-52-1161　Fax:0722-59-3340
1-1 Gakuen Sakai, Osaka 593 Japan

２．１ 日本の検索とＷｅｂ検索

「日本の検索」コーナーでは日本語の検索・索引サービスを紹介し、各サービスごとに、特徴、入力条件、検索結果、ニュースなどを詳しく紹介している。「Ｗｅｂ検索」では主に英語の検索・索引サービスを紹介している。リンクを張るだけでなく、メタ検索というか、検索キーを１回入力するだけで複数の検索サービスを検索できるようにした。これはJavaScript 技術で実現しており、フォームを使っているほとんどの検索画面に適用可能である。検索はビギナーの人も容易に行え、検索の普及に貢献できればと願っている。

２．２ オール索引と新着情報集

オール索引は各索引サービスの大分類項目や中分類項目を収集整理し、五十音別に配置した索引の索引である。索引前後の項目名も一緒に表示されるため、各索引サービスの分類項目の特徴や知識の構造が理解できる。何を探したらよいか分からない人や網羅的に情報を集める人が利用できる。さらに、新着情報集を利用すると、各索引サービスで収集され公開される新着情報のアクセスが容易になる。新着情報はデータベース化の遅れをカバーするため、あるいは新規サイトの発見のためにも必要である。

２．３ 内外の検索ニュース

今年に入ってからでも、規模の大きい新規の３サービスの参入があった。また、大改装した所は４サービスあり、その他の所も何らかの改装をしている。索引系の場合は、情報量の増加で分類項目を分割、統合、新設などがある。また、ホームページの新装、検索の入力画面や出力画面の機能追加などがある。変更がアナウンスされるのはまれなので、１週間に２回モニターしニュースとして流している。海外の場合は、主に海外のニュースをベースにしているが、すべてをフォローできていない。

２．４ 検索の視点、検索調査

検索の視点では、Ｗｅｂ情報の特徴や検索を取りまく環境などを、現実のデータをもとに分析し報告している。検索調査は４種類の検索キーを使い、各検索サービスごとの検索数を求めてテーブルにして発表している。検索サービスごとに検索手法が異なるため、検索の際に問題になった点を指摘し、提案もしている。さらに検索数だけによるランキングも行っている。提供側とユーザー側の検索技術の向上に役立てばと思っている。

２．５ ブラウザ、JavaScript

WWW情報はブラウザを使って表示するため、ブラウザの知識は必須である。そこでブラウザの話題をあつかうコーナーを設けた。現在、ブラウザのダウンロードの記録が中心になっているが、それを見てもブラウザのバージョンアップの激しいことがわかる。さらに、JavaScript のコーナーを設け、情報発信の際の HTML に利用できる小さなスクリプトを中心にあつかっている。

２．６ 掲載 Thanks、Link Thanks、ヒストリー

情報発信する際の参考になればと思い、各 Thanks コーナーとヒストリーをつくった。わずか３カ月半で、Link Thanks が 97 になり、掲載分を含めれば検索デスクを紹介している所が 100 以上ある。主に、検索キー asaisan を使って Alta Vista から検索した。長い間 Referation Analysis の研究をしていたが、ハイパーリンクの関係からその逆リンクが即座にしかも無料で得られるインターネットの世界には驚嘆する。

3．日本の検索・索引サービス

インターネットの世界では、多種多様な情報が飛び交っている。その中でも検索や索引サービスとその紹介は情報以外の専門分野の人、情報に関心のある人、学生、出版社やソフト会社など、情報専門家の手を離れて行われているといっても過言ではない。

主要な検索・索引サービスを表1に示した。検索主体が8、索引主体が12である。その運営は、大学の学生、ソフト会社、会社のソフト部門、出版社、会社、個人などさまざまである。まだ、その多くは無料奉仕のボランティアであり、徐々にではあるが利益を目標にする企業へと移行しつつある。

1年前に企業化を果たした米国の検索サービスが最近相次いで店頭市場へ上場し豊富な資金を手にしている。その資金と検索技術をもって、日本へ進出あるいは計画している。仲良しクラブで発展していく日本の社会の仕組みも大きな転換期を迎えている。Ｗｅｂ情報の急増、Ｗｅｂ情報表現の多様化など検索・索引サービスが抱える問題は多い。次節でそれらについて考察する。

第1表　日本の検索・索引サービス一覧

検索主体のサービス	運　営	内容	収　集	総数
ＯＤＩＮ	東大:原田昌紀	検索	ﾛﾎﾞｯﾄ	189,160
ＴＩＴＡＮ	NTT:関西支社	検索	ﾛﾎﾞｯﾄ	#300,000
Ｍｏｎｄｏｕ	京大: 情報通信	検索	ﾛﾎﾞｯﾄ	*200,000
千里眼	早大:田村健人	検索	ﾛﾎﾞｯﾄ	#97,000
InfoNavigator	富士通	検索	登録,ﾛﾎﾞｯﾄ	?
WAVE Search	ソニー	検索	登録	?
WWWナビゲーター	ｲﾝﾌﾟﾚｽ	検索	登録	13,189
Nippon Search Engine	マジックマウス	検索	登録,ﾛﾎﾞｯﾄ	*145,000
索引主体のサービス	**運営**	**内容**	**収集**	**総数**
NTT Directory	ＮＴＴ	索引,検索	登録	#5,200
Hole-in-One	日立国際ﾋﾞｼﾞﾈｽ	索引,検索	登録	*13,625
Japan Search Engine	京都:学生	索引,検索	登録	#5,000
Yahoo! Japan	日本ヤフー社	索引,検索	登録,ﾛﾎﾞｯﾄ	*15,526
CSJインデックス	ｻｲﾊﾞｰｽﾍﾟｰｽJ	索引,検索	登録	9,765
Yahho	tutkie:甲斐大樹	索引,検索	登録	11,128
URL広場	ORIONS	索引,検索	登録	*3,003
NETPLAZA	ＮＥＣ	索引,検索	登録,ﾛﾎﾞｯﾄ	5,515
日本ネット	日本ﾈｯﾄ研究会	索引,検索	登録	?
ＪＯＹ	斎藤正紀	索引	登録	3,745
日本企業URLﾃﾞｨﾚｸﾄﾘｰ	日経BP社	索引	登録	2,033
WWWファインダー	毎日新聞社	索引	登録	#1,160

注）総数は5月10日時点のもので、＊は最新の推定値、＃は古いデータ

4．日本の現状と課題

4．1 日本語Ｗｅｂ情報量の急増

誰も管理者がいないインターネットの世界で、日本語のＷｅｂ情報がどれ位あるかは誰にも分からない。索引サービスの中には登録情報を公表している所があり、それらを毎週統計データとして収集している。その一部を第2表に示した。

これらのデータは自薦で登録した情報であり、多くの索引サービスのデータ源である。今年の1月12日から5月10日迄の4ヶ月間で約2倍強になっている。これは処理する情報量が8ヶ月で4倍、1年で8倍になることを示す。勿論、月日が経るにしたがい鈍化していくとしても、3〜4年で100倍に達するとみなせる。この情報量の急増は処理システムの増強、回線の大容量化などの問題を提起している。

第2表　日本語Ｗｅｂ情報の推移(増加率は1/12から5/10)

サービス名	96. 1. 12	96. 2. 9	96. 3. 8	96. 4. 12	96. 5. 10	増加率
WWWﾅﾋﾞｹﾞｰﾀ	6, 460	7, 539	8, 901	10, 991	13, 189	2. 042
Yahho	5, 173	6, 408	8, 010	9, 897	11, 128	2. 151
NETPLAZA	2, 197	2, 741	3, 283	4, 548	5, 515	2. 510
ＪＯＹ	1, 724	2, 243	2, 625	3, 357	3, 745	2. 172
日本企業URL	1, 163	1, 377	1, 554	1, 813	2, 033	1. 748

4．2 索引系から検索系へ

日本の検索・索引サービスは米国Yahoo!を見本にしており、ホームページ製作者からの自薦情報の登録に依存している。しかし、窓口が一個所であれば網羅的な情報が集まるが、窓口が多いため情報が分散する。従って、網羅的な情報が収集できないため、複数のサービスを利用しなければならなくなる。

この4月にオープンした日本Yahoo!は登録とロボットを併用し、わずか2〜3ヶ月の準備でトップクラスの情報量で出発した。これによりこれまでの登録だけをベースにしたシステムからロボット併用型へ移行する兆しがでており、好ましい事態と言える。

日本の検索主体のサービスでロボットに依存している例は少なく、しかも大学の運用が多い。数年で100倍になる情報量の増加に対応できるだけの投資は大学では不可能であり、制約の少ない企業とはとても太刀打ちできない。米国の場合でも、どのような検索システムも1年も続かなく、何回も脱皮しながら存続を図ってきた。第1表のデータも年末を待たずに大きく塗り変わることが予想される。

4．3 最新情報の提供

索引主体のサービスはほとんどの所が週間単位で更新を行ない、新規に登録された情報は新着情報として公開している。この索引の場合は最新情報が入っているものとみなせる。しかし、ロボット依存型の検索システムでは最新情報の提供はなかなか難しい。検索調査で各サービスの検索数を調べているが、どの時点までのデータが入っているのか大体推測できる。それは最新のキーワードをどれ位検索するかで判断できる。

検索サービスによっては検索結果にデータ収集日や何KBの容量かを付加して表示する。これはユーザーにはとって非常に有用な情報である。古いデータからの検索結果しか得られないとしたら、探す情報にもよるがあまり役立たない。特に、4ヶ月で倍増する世界で、

4ヶ月も更新がないということは、旧い方の半分からの検索ということになる。

米国では最新情報の提供が一つの競争要因になりつつある。最近のAlta Vistaの検索では3〜4日前のデータが検索されることがある。ロボットの巡回アルゴリズムやデータベース化のスピードが驚くほど改善している。また、6月からサービス開始の infoSeek 社のUltraSeekも最新情報の提供を狙ったものである。

4.4 ブラウザ技術の進歩への対応

ブラウザ技術が急速に進歩している。これは歓迎すべきことであるが、この新しい技術から得られるＷｅｂ情報を処理する検索エンジンは十分に対応できない恐れがある。ブラウザ自体も、自らが提供した技術に対応するまでに数ステップのバージョンアップを必要とする。重要な情報が新しい技術を使った所に存在する場合には、それを見落とす検索システムは欠陥システムとみなせる。

昨年11月からフレーム機能が使えるようになった。しかし、フレーム表示の情報を処理できる検索エンジンは皆無に近い。フレームは一画面を複数画面に分割して、必要に応じて、その一部分を差し替えて情報を表示する。しかし、各フレームはそれぞれファイルになっており、それらを合成して一つの情報とみなしていない。例えば、目次が左側にあるような場合、ロボットにとって重要なリンク情報を収集できず、ロボットが素通りしていく可能性が高い。

これから本格的なマルチメディアの時代が訪れるが、検索はテキストデータを対象にしており、ShockwaveやVRMLの情報は扱わないというわけにはいかない。マルチメディア情報を処理する検索エンジンの開発も視野に入れる必要がある。

5. おわりに

日本の検索サービスもこれから大きく変化しようとしていた所へ米国の検索サービスが入ってきた。日本ヤフーの進出はボランティアの時代から企業競争の時代への引き金になり、今後一層加速する。これはユーザーにとっては良いことであるが、サービスする側にとっては大きな負担となる。検索に限らずＷｅｂの世界は試行錯誤をしながら段階的に発展するものであるが、ポテンシャルの高い競争相手の出現で180度の転換が必要となる。

インターネットの話題は盛んであるが、検索関連の投資はほとんどなされていないように思われる。企業のイントラも進行し、ますますＷｅｂ情報の処理が行われ、検索技術の重要性が増してくる。この新しい事態に無策というわけにはいかない。

最後に、Ｗｅｂ情報が急増している現状で一番大切なことは時間であり、時間を短縮するための方策を如何に考え実行するかにかかっているように思われる。

参照文献

[1] 浅井勇夫、ネット情報の一元管理用ソフトの開発：文書情報のHTML文書化と管理、第25回ドクメンテーション・シンポジウム予稿集、55-60(1995)

[2] 検索デスク、URL: http://www.bekkoame.or.jp/~asaisan/

[3] INTERNET Watch、特集：検索エンジン、1996年1月23日号、インプレス

ISSN 0387-7221

第26回 ドクメンテーション・シンポジウム

予稿集

Preprints of the 26th Documentation Symposium

期日：1996年6月20日(木)，21日(金)

会場：機　械　振　興　会　館
(〒105 東京都港区芝公園3－5－8)
Kikai-sinkō Building
(5-8, Sibakōen 3-Chome, Minato-ku, Tokyo, Japan)

主催　社団法人　情報科学技術協会
INFOSTA

後援
日本科学技術情報センター
社団法人　日本図書館協会
専門図書館協議会
日本データベース協会
日本医学図書館協会