ユーザの利用にあわせて成長するサーチエンジンの構築

林 賢紀
農林水産省農林水産技術会議事務局筑波事務所研究情報課
(農林水産研究情報センター)
〒305-8601 茨城県つくば市観音台 2-1-2
Tel: 0298-38-7285
FAX: 0298-38-7364
E-Mail: tzhaya@affrc.go.jp

概要

既存の検索エンジンでもつ検索機能に加え「ユーザの利用にあわせて収録データを拡大し自分のデータベースを成長」する機能を実装したサーチエンジンを開発した。このシステムでは、検索結果となったページにアクセスした際にそのページの URL を保存し、この URL とそのページから一階層のリンクをたどりページを取得し、データベースを再構築している。これにより、検索エンジン上のデータをよりノイズの少ないものに保ち、WWW 上の情報検索をより効率的に行うことを目的としている。

1998 年 9 月から 10 月までの間、国内の図書館の Web ページをデータベース化し、図書館職員を対象に試行を行った。多用なページの収集によりデータベースも拡充され、結果としてヒットするキーワードも増加する結果が得られた。

キーワード

インターネット、サーチエンジン、WWW、図書館

Making of the Auto Grown search engine and tryal.

Takanori HAYASHI
Research Information Division, Tsukuba Office, Agriculture, Forestry and Fisheries Research Council Secretriat. Ministry of Agriculture, Forestry and Fisheries.
(Agriculture, Forestry and Fisheries Research Information Center)

Abstract

This archicle tells about making of the Auto Grown search engine and tryal this system. This system update database by users activites. The auther try to develop the search engine which implemented a function "Evaluation of users to collecting data" in addition to search function had with existing search engine. By this system, the auther store URL of the page in the case that accessed the page which became search results and I trace one class of link from this URL and the page and acquire a page, and restructuring this database. And the result was get many pages and updated database.

keywords

internet, search engine, World Wide Web, Library

1. はじめに

インターネットという情報の発信個所が各所に分散した環境において、WWW は情報の形態やプロトコルの異なる個々のサービス同士をハイパーリンクで結合し、容易にアクセスできるようにしたシステムといえる。既存のサービスと異なり簡易な手順で利用でき、また情報の発信も容易にできることから爆発的に普及し現在のインターネットの急速な発展の原因の一つとなった。

しかし、発信の容易さはインターネット上の情報を劇的に増加させた。WWW では個々のサイトから情報を発信でき、情報同士をハイパーリンクにより容易に結合できる反面、一個所に情報の所在や概要を集積するようなシステムとしては設計されていなかったため、WWW サーバの増加は同時に情報の洪水を生んだ。結果、情報の所在や内容を把握しにくくし、有益な情報が埋もれてしまうことになる。

このため、インターネット上の情報を整理して提供するサービスが自然発生的に生まれた。比較的古くから存在するのは、The WWW Virtual Library[1] と呼ばれる WWW での情報を項目別に分類しディレクトリ形式でまとめたもので、これは世界各地で分野を分担して作成されている。同様のディレクトリサービスの一つ Yahoo!では、ホームページを開設したユーザからの登録申請を元に分野ごとの複数の担当者が登録作業を行っている。

一般に「検索エンジン」あるいは「サーチエンジン」などとも呼ばれる、自動的に WWW 上の情報を集積しデータベースにするサービスも行われている。多くの場合、「スパイダー」あるいは「ロボット」と呼ばれる自動検索プログラムが日夜全世界の WWW サーバを巡回して情報を集め、データベースを作成している。日本国内では goo や infoseek、altavista などのサービスが有名で、多くのユーザに使われている。これらのサービスは非常に多くの Web ページを収録しており、網羅的な検索を行う場合は有効であるが、網羅的であるがゆえに検索結果中には必要としない情報（ノイズ）も多い。また、サービスごとに使用できる検索機能や網羅性に特徴があるため、場合によっては使い分けることも必要となる。

これら既存の検索サービスの現状の問題点としては、

- 収録範囲が多いため、検索結果にノイズがある
- 収録範囲の増大とともに、サービスの維持管理のコストも増大する

などが挙げられる。

そこで、これらの問題点を解決し効率的な情報検索を行いうるシステムとして、以下の方法をとる検索システムを開発した。

- 専門分野に特化したデータベース
- 収録するデータは利用者によって評価し、有効なデータのみを収録

まず、データベースの専門化であるが、収録する分野を決定しデータを収集するサイトを特定することで、ある程度の収録データの専門化が図れると想定している。また、検索結果の表示時に提示されたサマリーを参照し、この中から有効と思われる URL へアクセスすることで評価をしたものとする。

以下、このような方法を持った検索システムの概要や試行の状況などについて述べる。

2. システムの概要

このシステムでは、既存の検索エンジンでもつ検索機能に加え「ユーザの利用にあわせて収録データを拡大」する機能を実装している。

検索エンジンは、多くの場合は自分の必要とする情報を検索するために利用され、利用者は検索した結果のサマリーを見てアクセスするかどうか判定している。このシステムでは、このような利用者の行動をもってユーザによるデータの評価としており、検索結果となったページにアクセスした場合はそのページの URL を保存している。そして、この URL とそのページから一階層のリンクをたどったページを取得し、データベースを再構築している。このような手順により、ユーザは特に意識することなくデータベースに追加するデータを決定することができ、評価されたデータのみを採録できると考えられる。

データベースシステムには Namazu[2] を採用している。Namazu は高林によって開発された日本語全文検索システムで、GPL2（GNU 一般公有使用許諾書バージョン 2）に基いたフリーウェアである。手軽に使えることを第一に目指したシステムで、CGI(Common Gateway Interface) として WWW 上のデータベースとして動作するほか、自分の所有するパソコン上のファイルを対象としたパーソナルなデータベースの構築にも利用できる。

Namazu の特徴としては、

- ソースが公開されている
- ユーザのメーリングリストで稼動状況の報告や情報交換が行われ、機能拡張が迅速に行われている。
- 高速かつ導入が手軽

などがあるが、今回 Namazu をデータベースとして採用した理由は「手軽に導入と運用ができ、比較的高性能」な点による。また、ソースが公開されているため、実行時の挙動がつかみやすく柔軟な運用が可能なことも評価している。さらに、kakasi[3] を用いて語句の分かち書きを行うが、辞書を拡張することによりインデックス時の精度を向上できるなど、専門分野に特化したデータベースの作成に有効だと思われた。

データの取得には wget[4] を利用している。wget は GNU によるフリーウェアで、WWW サーバ上に置かれた robots.txt を読み取りデータ取得の可否を判定するため、情報提供者側の意思を多少なりとも反映できるものと考えられる。

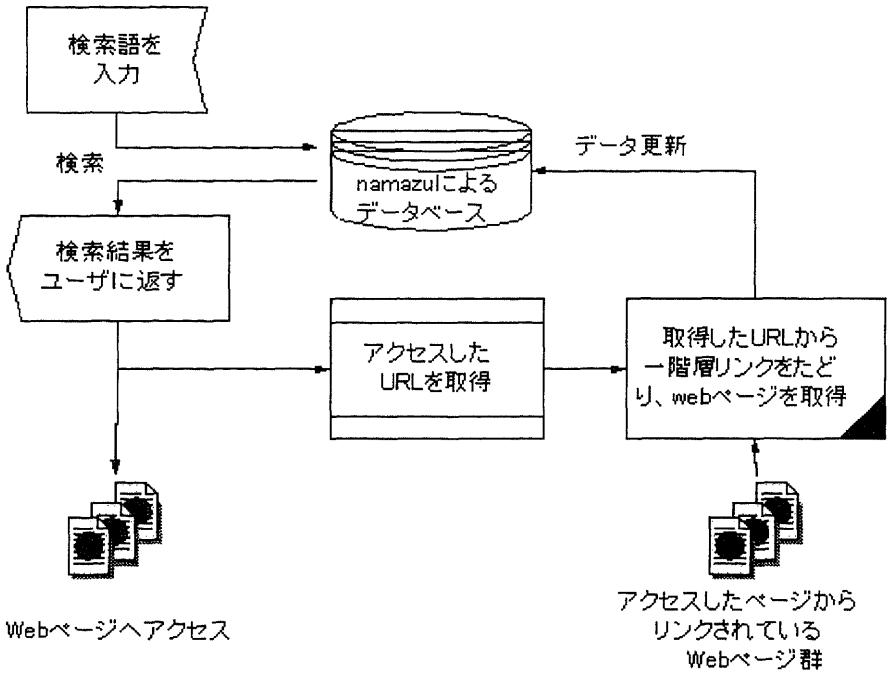
ユーザがアクセスしたリンクの記録は、oidon が作成した Link Checker[5] を改造し利用している。もともとは、アクセス回数などを記録する perl で記述されたスクリプトであったが、今回はこれに手を加えて取得対象となる URL を別途記録するように改造を行った。

これらを組み合わせ、シェルスクリプトを用いて定期的にページを取得しデータベースの再構築を行うシステムを構築した。図 1 に、本システム「自動成長型サーチエンジン」の概要を示す。

3. データベースの構築

今回開発したシステムを利用し、国内の図書館で公開されている Web ページを収集しデータベースとすることにした。すでに、国内の図書館で運用されている WWW サーバ約 100 個所を収録対象とした検索サービス「国内図書館 web 検索システム」[6] を試行的に運用していたが、サイトを巡回してのデータの収集と更新に時間を要するなど、逐次の更新を行うためには問題が多かった。特定多数のサイトからロボットで Web ページを取得するが、リンクをたどりすべてのファイルに対して定期的にアクセスするため、サーバへの負荷やネットワークのトラフィックの増大などが危惧された。

図1　自動成長型サーチエンジン
システム概念図



そこで、今回の試行では、林が作成・メンテナンスをしている国内の図書館で目録サービスを提供しているサイトを集めたリンク集「Jump to Library! (in Japan)」[7] を基点とし、ここから一階層の Web ページを取得し最初のデータベースとした。これにより、国内の 100 個所近い図書館のホームページと OPAC(オンライン利用者目録) のページが最初のデータベースとなった。これらのデータの集合に対しユーザが検索を行い、リンクされたページにアクセスすることで、このデータベース自体がさらに成長して行くことになる。収集する分野を図書館関連のページとしたのは、利用者としてインターネットでの情報検索について興味や経験のあるユーザ層である図書館職員を想定し、かつ収集する情報に対してある程度の評価ができるであろう点が理由である。

Web ページ取得は、対象となる URL を蓄積しこれをまとめて cron により毎晩午前 1 時ごろ自動的に行っている。データベースの更新は、取得したページの数によるがおおむね 2 時間以内には完了している。このとき、対象分野以外のページはなるべく取得しないよう yahoo や infoseek, NTT など大規模なディレクトリを提供しているページは、取得対象から除外している。

参考のため、今回のシステムが稼動している環境を以下に示す。

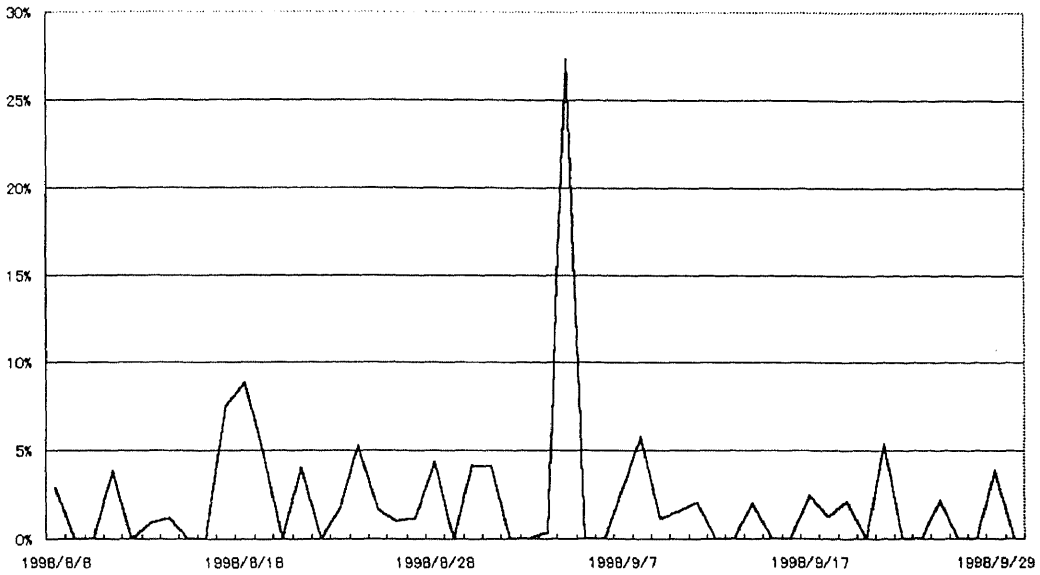
```
マシン：Digital AlphaServer 2100 4/275
メモリ：2 Gb
OS　　：Digital UNIX 3.2G
```

4. 利用状況と考察

1998年8月から、このシステムを大学図書館職員などで構成されるメーリングリストで広報し、試行を開始している。

データベース自体も徐々に成長しており、さまざまなWebページを収集している。以下、いくつかの表にて1998年8月から9月までの利用の状況を示す。表1は、ヒットしたページにアクセスした割合であるが、検索結果全体のURLから見ると数%程度であることが分かった。これは、

表1 ヒットしたページにアクセスした割合



- 検索はしたものの実際には検索結果にアクセスはしなかった

あるいは、

- いくつかの検索結果の中からひとつを選択して閲覧し、そこからリンクをたどって行くことで目的のページにたどり着いた

かのいずれかの理由が考えられる。表2は、実際に検索語として使われたキーワードが、データベースの成長の過程でどのようにヒット数が変化したかを示す。いくつかのキーワードでは、利用に応じてヒット数が増加していることが分かる。表3は、取得したページと、インデックスされたページとの比較である。これは、データベースがどこまで成長するかを表しているが、約2ヶ月という期間では取得したページとインデックスしたページのとの比率に特段の変化は見られない。このことから、現時点では利用に応じた一定の割合で成長しているものと考えられる。また、表4が実際にアクセスされたページのURLとその回数である。全体では245のURLがアクセスされているが、その半数以上は1回程度のアクセスにとどまっている。5回以上のアクセスのあったWebページは6URLに過ぎなかった。表5はアクセスの多かったページであるが，インターネット上のリソースへのリンクをまとめたページが多い。

以上の点から、今回の約2ヶ月間の試行に於けるユーザの利用とその結果をまとめると、ユーザがアクセスするページは特定されず、このため収集対象とするページも多様化している。また、多様なページの収集

によりデータベースも拡充され、結果としてヒットするキーワードも増加する結果となった。さらに，アクセスされたページを見ると，他所へのリンクをしているページは参照されやすいことが考えられる。

5. 関連の研究

今回のアプローチ以外にユーザが利用したデータを収集する手法としては、proxy やクライアントマシン上に蓄積される cache を利用するという方法[8][9]がある。通常、cache は一度取得したデータを複数で利用するなどして効率的に利用しアクセスの高速化やネットワークの効率化を図る手段であるが、一度アクセスしたページをデータベース化し共有することで、今までユーザ一人のものであった情報や検索手法を共有する、というものである。

特に、清水らの PA-search[8] では、実際にアクセスしたデータを共有することで検索需要のうちおよそ2割を支援することに成功するなど、小規模な実験ながら妥当な結果を得ている。これらの手法を応用することで、より高い精度でユーザの検索行動を支援することができるのではないかと考えている。

6. 今後の課題と機能拡張

本稿では、1998 年 8 月から 9 月までの試行を通じての結果について取り上げたが、今後もさらに運用を行うことで、長期にわたる利用状況などを把握し検証を行うことを考えている。

今後の課題としては、ユーザの検索行動の把握が挙げられる。たとえば、検索結果からアクセスされた Web ページが、ユーザの希望に合致するかどうかを確認することもデータの評価という点では必要であろう。

また、データの精度をさらに向上させることも必要であると考えられる。10 月からは、筑波大学図書館部の近藤が作成した図書館用かな漢字変換辞書パッケージ l-dic[10] を kakasi に組み込み、図書館用語などをインデックスとして切り出している。これにより、図書館職員がよく使う用語での検索の際にヒット率が向上するものと想定される。また、新規に取得した Web ページを別にデータベース化し検索できる機能を追加した。現在、この機能を使い新規に取得した Web ページの中からユーザが指定したキーワードで検索を行い、一定以上のスコアを記録したページのみをデータベースに追加する機能の実装を進めている。将来的には、ユーザの指定した語に加え、類義語辞書を利用して語彙を拡張した上でフィルタリングを行ってデータベースに追加することを考えている。これに加え、最初にデータベースとした Web ページの質や特性によって、収集されうるページも変化することが予測される。今後は、データ収集の基点となる Web ページの選定や評価の手法についても検証を行いたい。

今後、本システムについては、研究者のグループや個人など小規模な集団での利用を考慮して構築を行い、より収録分野を絞った検索エンジン作成用システムとしていきたい。具体的には、研究者など専門知識と情報検索を行いうるグループよりその分野の基礎となる Web ページを聞き取った上で、そこからいくつかのページを収集しデータベースとする。そして、このデータベースを利用してもらうことで、グループにより評価されたページを収集して行く。いわばその道のプロによる「データベースの調教」を行い、データベースを適切な形で拡張するのである。これをいくつかの分野について行い、最終的には個別の専門分野が統合された検索エンジンの構築を目指すものである。

参考文献

[1] The WWW Virtual Library:http://vlib.stanford.edu/Overview.html,1998 年 10 月 1 日アクセス

[2] 高林哲、日本語全文検索システム Namazu:http://saturn.aichi-u.ac.jp/~ccsatoru/Namazu/intro.html.ja, 1998 年 10 月 1 日アクセス

Namazu でのデータベース構築と利用については、馬場肇, 日本語全文検索エンジンの構築と活用, ソフトバンク. 1998.9 に詳しい。

[3] 高橋裕信, ftp://ftp.kusastro.kyoto-u.ac.jp/pub/baba/wais/, 1998 年 10 月 1 日アクセス

kakasi に、馬場肇による分かち書きパッチを当てたものを namazu では使用する

[4] Hrvoje Niksic 、GNU wget:http://sunsite.auc.uk/ftp/pub/infosystems/wget/, 1998 年 10 月 1 日アクセス

[5] oidon, Link Checker:http://www.iod.co.jp/~oidon/,1998 年 10 月 1 日アクセス

[6] 林賢紀、国内図書館 web 検索システム:http://www.affrc.go.jp/~tzhaya/library/seek4lib.cgi, 1998 年 10 月 1 日アクセス

本稿での内容を含めこのシステム全体の概要については、ttp://www.affrc.go.jp/~tzhaya/library/ ,1998 年 10 月 1 日アクセスを参照されたい。

[7] 林賢紀、Jump to Library! in Japan:http://ss.cc.affrc.go.jp/ric/opac/opac.html, 1998 年 10 月 1 日アクセス

[8] 清水奨, 神林隆, 佐藤進也, その他、グループ試行 WWW 検索アシスタント PA-search の実現: http://www.ingrid.org/w3conf-japan/97/shimizu/pas-info.html, 1998 年 10 月 1 日アクセス

[9] 吉岡恒夫、代理サーバを利用した検索システム:http://infonet.aist-nara.ac.jp/member/tsuneo-y/, 1998 年 10 月 1 日アクセス

[10] 近藤努、図書館用かな漢字変換辞書パッケージ l-dic:http://www.tulips.tsukuba.ac.jp/~kondou/ldic/, 1998 年 10 月 1 日アクセス

表2 キーワードとヒット数の増加

検索日時	キーワード	ヒット件数
Tue Aug 11 12:40:16 1998	java	19
Thu Aug 13 08:44:39 1998	java	19
Wed Aug 19 15:12:13 1998	java	53
Fri Aug 14 15:01:56 1998	webcat	70
Thu Sep 10 13:14:11 1998	webcat	78
Mon Aug 10 19:25:45 1998	z39.50	3
Mon Aug 31 20:38:41 1998	z39.50	5
Thu Sep 10 13:16:42 1998	z39.50	7
Thu Aug 13 13:31:35 1998	国文 学研 究	13
Tue Aug 18 11:42:31 1998	国文 学研 究	13
Wed Aug 26 15:22:09 1998	国文 学研 究	14
Thu Sep 3 11:12:02 1998	国文 学研 究	15
Fri Aug 14 15:16:26 1998	中国語	27
Wed Aug 19 11:58:32 1998	中国語	29
Thu Sep 17 10:59:28 1998	中国語	66
Fri Aug 14 15:03:48 1998	電子図書館	66
Thu Sep 10 17:25:18 1998	電子図書館	74
Thu Sep 17 10:55:55 1998	電子図書館	80
Sun Aug 16 22:31:18 1998	農業	40
Wed Aug 19 12:28:18 1998	農業	48
Fri Aug 21 11:47:25 1998	農業	49
Tue Aug 25 21:39:34 1998	農業	60
Mon Aug 31 15:47:03 1998	農業	65
Wed Sep 9 23:35:02 1998	農業	69
Mon Sep 14 19:39:40 1998	農業	71
Fri Sep 18 14:12:38 1998	農業	80
Sat Sep 19 01:21:17 1998	農業	94
Sun Sep 20 22:23:05 1998	農業	97
Tue Sep 22 08:52:36 1998	農業	108
Mon Sep 28 15:50:11 1998	農業	108
Wed Aug 19 11:52:57 1998	文字コード	6
Thu Sep 10 13:27:32 1998	文字コード	19
Fri Sep 11 13:00:26 1998	文字コード	22
Thu Sep 17 10:51:34 1998	文字コード	32
Mon Aug 17 13:12:48 1998	幕末	16
Wed Aug 19 10:40:14 1998	幕末	28

表3 取得したページとインデックスしたページ

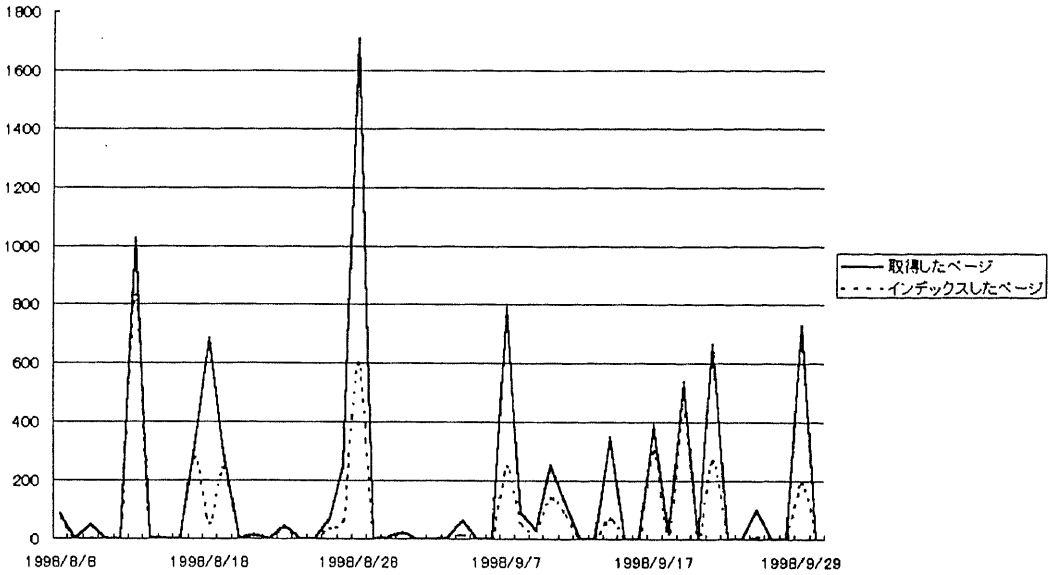


表4 URLごとのアクセス回数

	URL
11回のアクセス	1
9回アクセス	1
8回アクセス	0
7回アクセス	2
6回アクセス	1
5回アクセス	1
4回アクセス	10
3回アクセス	12
2回アクセス	37
1回アクセス	180
合計	245

表5　アクセスの多かったページ

URL	アクセス数
http://www.ntt.co.jp/SQUARE/www-in-JP-j.html 日本のWWWサーバ（NTT）	11
http://www2.ll.chiba-u.ac.jp/~yamamoto/link.html Internet Resource Selection（千葉大学）	9
http://www.n-lib.toho-u.ac.jp/zasshi/kenkyu.htm 研究室資料について（東邦大学習志野図書館）	7
http://www.tulips.tsukuba.ac.jp/other/japan.html 日本の図書館（筑波大学）	7
http://www.lib.ynu.ac.jp/LINK/library.html 本と図書館に関する情報（横浜国立大学）	6

アクセスの多かったページのうち、上位5つ