コンピュータと教育 59－5
（2001．3．23）

BioMetasearch

－生物資源系メタ検索インターフェイスの実装－

古瀬慶博，林　克之，吉野直子，目黒也智#1，木下哲男#2，木原英人#3，鵜川義弘#4

汎用のサーチエンジンサイトを利用した情報検索は，その性格上情報収集能力に限界がある．生物資源系サイトは，コンテンツおよびデータ構造(リンク空間）の update が頻繁なため，ソフトロボットによる URL 収集の利用価値は必ずしも高いと言えない．また生物資源系サイトは，ソフトロボットでは収集できないサーバも数多く存在する．これらの問題を解決するために，Metasearch をエンジンとした生物資源系サイトを対象にしたメタ検索システム BioMetasearch を構築した．BioMetasearch は，一回の問い合わせをもとに，汎用の検索サイトと生物資源系の検索サイトの複数に対して同時に検索を行い，結果を統合して表示することができる．

BioMetasearch

NOBUHIRO FURUSE, KATSUYUKI HAYASHI, NAOKO YOSHINO, NARITOMO MEGURO#1,
TETSUO KINOSHITA#2 , HIDETO KIHARA #3 and YOSHIHIRO UGAWA#4

Information retrieval by using sites of the general-purpose search engine has inclusively constrain those of URL collection ability. The biological web sites do update thier content frequnetly, so that their utility value of the URL collection by the soft robot comes to be not always high. Also there are in great numbers on the biological web server whom it can not collect in soft robots. We construct new retrieving interface for biologist, called BioMetasearch. Our new retrieval interface for extracting bioresource on the Internet is running on the Metasearch platform, which have being developed by Tohoku University. For both the multiple of general-purpose retrieval sites and usual bioresource gateway, the BioMetasearch can obtain the result of integrating in a query of one time.

1. はじめに

FTP, Gopher SpaceからWWWへとインターネットにおける情報空間の記述方法の主流は変革を遂げてきた．1993 年末に登場した Mosaic[1]は，それまでテキストベースででしなかったハイパーリンクの概念[2]を視覚化したブラウザであった．同時に，Mosaic の出現は，テキスト内でのリンク構造に自由度を与えた結果，情報空間全体の大きさ，言い換えると情報空間の地平線を測りにくくした．

ディレクトリ構造をもつ FTP や Gopher Space は，有向グラフ（探索木）で表現可能である．したがって，これらで構造化された情報空間の大きさは，測度可能であった．すなわち，ディレクトリの枝刈りによって，トップ（始点）とボトム（終点）が認識となる．一方，ハイパーリンクはディレクトリ構造からの制約から逃れるため，リンクされている全空間を探索できたかどうかが，保証されなくなる．加えて，URL 生存時間とよばれている，リンク構造が時間とともに書き換えられていく可能性を考えると，

#1 三菱スペース・ソフトウエア（株）
Mitsubishi Space Software Co.ltd. furuse@mi.mss.co.jp
#2 東北大学 電気通信研究所
Research Institute of Electrical Communication., Tohoku Univ.
#3 東北大学 電気通信研究所 （現：（株）富士通研究所）
Current: Fujitsu Labotatories.
#4 農林水産省農業生物資源研究所（現：宮城教育大学）
Current: Environmental Education Center., Miyagi University of Education ugawa@ipc.miyakyo-u.ac.jp

巡回型のソフトボットによる情報収集能力には限界があると言える．これはかなり以前から本質的な問題として認識されており，Northen Light[3]によると，巡回型のソフトボットを利用した検索エンジンサービスは，WWW コンテンツの全体に対して 16％程度のカバー率であることが指摘されている．このカバー率をあげるため分散協調型の収集も試みられている[4].

ソフトボットは，ロボット三原則に従い，ある規約[5]（たとえばウエッブサイトに置かれた robots.txt ファイルの参照）にそってウエッブ内を徘徊する．このため，ソフトボットにまったく不可視なサーバーも数多く存在することが指摘されている[6].

現在国内で公開されている検索エンジンは，Yahoo のような人手を介したイエローページを除いて，15 程度知られている[7]．またポータル用な商用全文テキスト検索エンジンは，20 程度存在する[8]．これらソフトボットの仕様，検索手法および検索結果のランキングの方法などは，かならずしも公開されていない．したがって，機能性能について，ベンチマークは不可能である．よって，ad hoc なウエッブ空間での情報収集でなく，散在する情報を網羅したい場合には，複数の検索サイトの利用は，不可欠といえる．

さらに，近年では，ASP[9],Namazu[10]で代表されるローカルエンジンの普及に伴い，明示的なリンクが行われないサイトの存在が指摘される．これらのウエッブサイトは，明示的な内部へのリンクをもたせなくてもよい．このことは，内容の更新が頻繁なサイトにおいては，リンク構造の変更が不要となる．反面，外部からの情報抽出には，キーワード（クエリー）を与えない限り，情報空間の全体は見えない構造となる．この種のウエッブサイトは，更新された URL の存在が外部から隠蔽されている．このため，ソフトボットでは容易に対応できない．

このような背景を踏まえながら，特定のキーワードに関連した情報を網羅的に得る手段として，複数の検索エンジンの利用に加えて，ローカルエンジンをもつサイトへ検索をおこなうことは必須と言える[11]．本研究は，生物資源系分野の情報を得るために，検索サイトへ一括して検索を行い，得られた情報を統合するシステム BioMetasearch の構築およびインターフェイスについて報告する．

2. 生物資源系サイトとメタ検索

検索エンジンへの検索は通常メタ検索とよばれている（図１）．海外では MetaCrawler[12]や Savvy[13]がメタ検索を実装している．

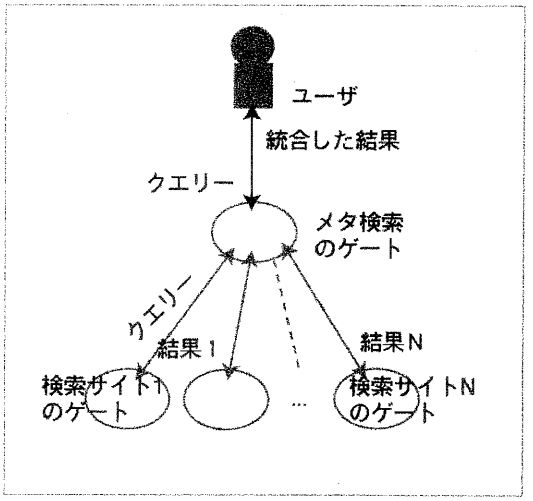

図１　メタ検索の概念

生物資源系とりわけゲノム情報に関する分野では，ウエッブで公開されているデータは時としてサイトあたり一日あたり数千ページ増加する場合がある．また，インターネット上でのデータ交換フォーマットには，ＸＸ方式と呼ばれるようなデファクトが存在する．これらの既存の流通フォーマットと矛盾しない範囲で文書内に記述されたハイパーリンクを辿ることで，分散したサイトから完全な最新情報を入手することが可能となる．

たとえば，ゲノム情報として，インターネットで 1 file を 1 directory でオンライン登録する古典的な管理がなされた場合，一日 1 万件の登録がなされたとき，directory と 10000 のウエッブページが作られる．仮に，それぞれの登録文書中に 10 カ所のハイパーリンクが存在するとすれば，新たなリンク情報だけでも 10 万

件に達する．これらのリンク先が同程度の頻度で更新されるとすれば，ソフトボットにより巡回することは事実上不可能といえる．

生物資源系のウエッブサイトは，検索のゲートをローカルに持つことでリンクの整合性の作業を軽減し，保守性を容易にしている．すなわち，検索キーワードを受け付けるたびに，動的にフラットなテキストファイルを抽出し，リンクのアンカーを生成する（たとえばSRS[14]）．

見方をかえると，このインターフェイスは，ユーザへの構造隠蔽を行っているといえる．すなわち，ユーザは問いかけ（クエリー）をしない限り，データを入手できないだけでなく，クエリーを送ることによりユーザ情報（嗜好）を開示していることになる（クエリーとコンテンツの暗黙の等価交換）．ソフトボットはクエリーを投げないため，このようなサイトの中身は検索エンジンからは不可視となる．

一方，生物資源系の情報を網羅しようとすれば，上述の特定分野のロボット不可視なサイトに加えて，通常の検索エンジンとの併用は欠くことができない．最近では，通常の検索エンジンでは，URLフィルターを加えることもできる．そこで生物資源系に限定した複数サイトへの一括検索を自動化させたシステムがBioMetasearchである（図2）．

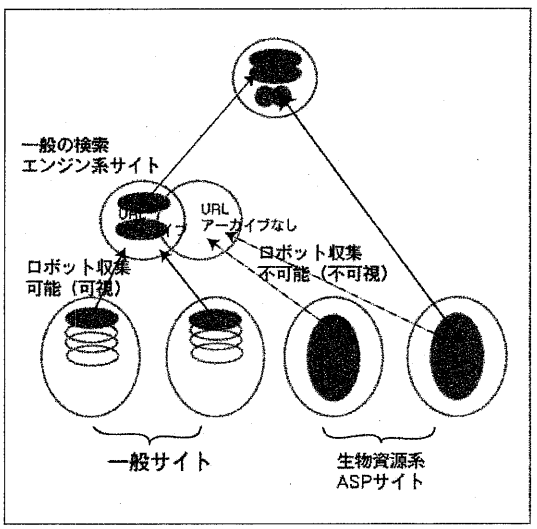

図2 BioMetasearchで目指す姿

（メタ検索ロボットがユーザクエリーに相当する）

3. 実装方法

3.1 機能概要

BioMetasearchは，主としてメタ検索エンジンmetasearchとページ収集プログラムsanshowowから構成される[15]．メタ検索エンジンは，ライブラリとしてeit-libcgiおよびw3c-libwww-5.1bを利用する．ページ収集プログラムは検索結果に同一URLがある場合に，重み付けにより重複したURLを削除する機能をもつ．

3.2 ユーザインターフェイス

実装しているインターフェイスを図3に示す．検索キーワード，URL指定，生物資源系URL,制限時間，汎用検索サイトの指定を1画面で行う．URL指定を行うと，汎用検索エンジンに対してキーワードとURLの論理積で検索を行う．たとえば，「DDBJ（国立遺伝学研究所）サーバーに存在する指定したキーワードのコンテンツだけを，検索せよ」といった内容を検索エンジンに対して依頼する．

図3： BioMetasearchの
ユーザインターフェイス[16]

検索エンジンがURL指定を受け付けるサイトとそうでないサイトとからの応答結果の統合までの流れ

を図４にまとめた.検索されたURLが複数ある場合には，どの検索エンジンでヒットしたかをユーザに明示的に知らせることが望ましい.

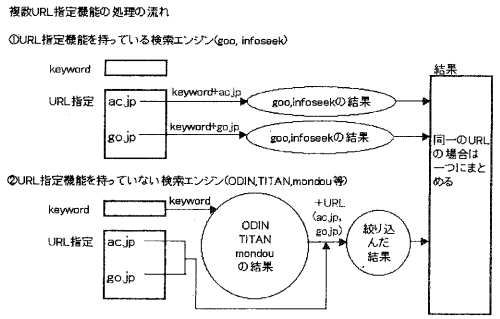

図４：キーワードとURL指定から検索結果統合までの流れ

4. 使用例

実例を示す．図５には，解読がまもなく完了されるとされている（99年10月段階）*Arabidopsis*（シロイヌナズナ）ゲノム[17]に関する情報を検索した結果である．検索対象は，検索エンジンとしてInfoseek,goo,google, 生物資源系サイトとしては GDB (The Genome Database) を指定した例である.

いくつかの実験を行うことにより，以下のような検索アプローチが戦略的に考えられる. 以下の事例は1999年３月時点のものである.

(a) 検索対象がどこのドメイン(URL)であるか既知である場合

指定方法：キーワードとURLを指定して検索

[例] キーワード：検索・エージェント・論文
URL：ai-www.aist-nara.ac.jp
検索エンジン： goo,infoseek
検索結果：11件

(b) URLが未知な場合

指定方法：キーワードを指定し，結果から再度検索

[例]手動による relevance feedback

（ステップ１）

キーワード：検索，エージェント，論文
URL：指定なし
検索エンジン：すべて
検索結果：243件

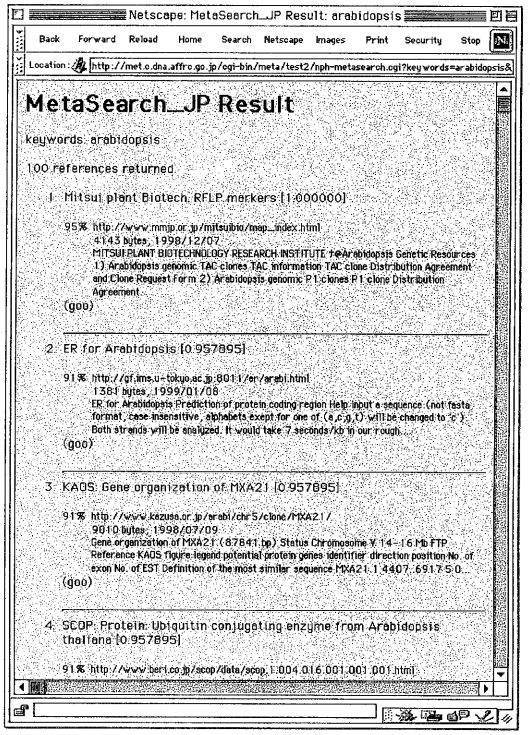

図５：キーワード *Arabidopsis* についてのメタ検索結果

>>検索結果から以下の点に着目

・大学，官庁研究所に情報が集中していないか (domain specific)

・ 特定のURLに集中していないか (URL specific)

・検索エンジンでヒット数がゼロのものを除外する

・検索結果のコンテンツ中に所望のキーワードがある場合はキーワード候補として絞り込

みに追加する

（ステップ２）

キーワード：検索，エージェント，論文

URL: ac.jp

検索エンジン：goo,infoseek,ODIN,TITAN

検索結果：201 件

(c) 特定のドメイン（たとえば大学のみ）

指定方法：ドメインで検索できる goo と infoseek で検索

[例] キーワード：検索，エージェント，論文

URL: ac.jp

検索エンジン：goo,infoseek

検索結果：99 件

使い方の詳細は，[16]にオンラインドキュメントを置いた．なお，2001 年 3 月現在で，オンラインドキュメントは公開しているものの、メタ検索の機能性格上、運用維持のためのボランタリーベースの手作業が追いつかない。このため，試験運用を休止している．今後は、環境研究者向けの Metasearch の構築を計画している．

5. まとめ

ソフトボット系の検索サイトおよびローカルで検索エンジンを実装する生物資源系サイトの両者に対して，一括で検索をおこなうことができるメタ検索 BioMetasearch の実装と使用例について示した．前者は汎用であり，後者は生物資源系に特化したサイトである．これらのまったく異なるレベルで情報検索および収集しているサイトに対して，研究目的に見合ったレベルで情報を網羅的に入手するというのは，極めて重要なインターフェイスでありながら，いくつかの不確定な要素の上に成り立っているといえる．

メタ検索は検索エンジンが存在しないと成り立たないだけでなく，検索エンジン側の受け取り方法が変わった場合に，BioMetasearch 側のインターフェイスを変更し，対応していく必要がある．さらには，昨今の i-mode による URL の自動抽出サービスのように，検索サイトから送られてくる cookie および広告の取り扱いについては，明確な指針と慎重な対応が要求されるであろう．通常の検索サイトとの共生関係を模索しつづけていきたい．

謝　辞

本研究は科学技術振興事業団平成 10,11 年度知的基盤整備推進精度「生物系研究資材のデータベース化及びネットワークシステム構築のための基盤的研究開発」の助成金の一部を使用した．また，農林水産省農業生物資源研究所，長村吉晃博士にご指導をいただいた．ここに感謝いたします．

文　献

[1]http://www.ncsa.uiuc.edu:80/SDG/Software/Mosaic/Docs/versions.html

[2] Lee,T.B., R. Cailliau, A. Luotonen, H. F. Nielsen and A. Secret(1994) The World Wide Web, Communications of the ACM,**37**,no.8,76-82

[3] Lawrence,S. and C.L.Giles (1998) Searching the World Wide Web, Science,**280**,98-100.

[4] 村岡洋一，田村健人，山名早人，河野浩之，森　英雄，浅井勇夫，西村英樹，楠本博之，篠田洋一 (1999) Internet 広域分散サーチロボットの研究開発，第 18 回 IPA 技術発表会論文集，71-78.

[5]Robot Exclusion http://info.webcrawler.com/mak/project/robots/exclusion.html

[6] 浅井勇夫 (1996) 検索可能性をあらわす検索力について http://www.searchdesk.com/view/vpt6916.htm

[7] http://www.ingrid.org/w3conf-bof/search.html

[8]全文検索システム協議会 (1999),全文検索システムとは何か？,FTSA,日経リサーチ情報開発局，pp.102

[9] http://www.aspnews.com/Pubs.htm など

[10]http://www.namazu.org/

[11]北村泰彦，野崎哲也，辰巳昭治 (1997),スクリプトに基づく WWW 情報統合支援システムとゲノムデータベース，ソフトウェアエージェントとその応用シンポジウム，電子情報通信学会，87-92.

[12] Seberg,E. and O. Etzioni(1995),Multi-engine search and comparison usng the MetaCrawler,World Wide Web Journal, 4th Int'l WWW Conf, Proc.,195-208, O'Relly & Associate Inc,. See http://www.metacrawler.com
[13]http://www.savvysearch.com
[14]http://www.embl-heidelberg.de/srs5/
[15] 木原英人，木下哲男，白鳥則郎(1998)，エージェントを用いたWWW情報検索システム，信学技報AI98-4(1998-05),電子情報通信学会,23-28,
http://www.shiratori.riec.tohoku.ac.jp/~kihara/metasearch.html
[16]http://bio-crawler.dna.affrc.go.jp/metasearch/
[17]http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/7227.html