Bull. Natl. Health Sci., **118**, 81-84 (2000)

分野に特化したインターネットの専用検索エンジンの開発

高井貴子，徳永雅彦[*1]，前田　憲[*2]，神沼二眞[#]

Development of Domain Specific Search Engines

Takako Takai, Masahiko Tokunaga[*1], Ken Maeda[*2] and Tsuguchika Kaminuma[#]

As cyber space exploding in a pace that nobody has ever imagined, it becomes very important to search cyber space efficiently and effectively. One solution to this problem is search engines. Already a lot of commercial search engines have been put on the market. However these search engines respond with such cumbersome results that domain specific experts can not tolerate. Using a dedicate hardware and a commercial software called OpenText, we have tried to develop several domain specific search engines. These engines are for our institute's Web contents, drugs, chemical safety, endocrine disruptors, and emergent response for chemical hazard. These engines have been on our Web site for testing.

Keywords: full-text search engine, information retrieval, domain specific information

はじめに

我々の部は，インターネットに接続される所内LAN環境，NICIを開発，整備する一環として，自らも情報を発信するWeb環境を整備してきた．これによって我々の部だけでなく，多くの部がWebページによる情報提供をはじめた．現在のところ，こうした情報提供は，各部門で自由に行われており，全体的な体系化，規制，統制はない．その結果，外部から見ると，どこにどのような情報があるかを知ることが難しくなってきた．そこで我々は，インターネットの検索エンジンと呼ばれるシステムを用いて，NICIの管理下にあるサーバー上におかれたすべての情報を一元的に検索するシステムを自ら開発し，1998年より試験的な利用に呈した．

その後このシステムを拡張して，インターネット上に提供されている多量の外部情報を自動的に収集し，インデックス化して，検索するシステムを試作した．だが使用実験でこのシステムは，ハードウェアもソフトウェアも十分でないことが判明した．そこで，1999年10月のNICI第3次システム整備において専用のマシンと商品ソフトを用意するとともに，平行して各種の専用検索エンジンを開発してきた．現在までに開発されているのは，内分泌かく乱物質関連情報用，医薬品関連情報全般用，医薬品緊急安全性情報用，GINC（Global Information Networks on Chemicals）用，健康危機管理関連情報用である．専用検索エンジンとは，専門分野の情報に特定した検索エンジンである．

本稿では，我々が行った検索エンジンのシステム開発について解説し，その有用性について述べる．

システム

最初に，検索エンジンについて簡単に説明する．検索エンジンは，指定されたサイト群から自動的にテキストを収集する「情報収集」と，集められたテキストをキーワードで検索する「全文検索」という，2つの基本的な機能を持っている．すなわち検索エンジンが持つ，インターネット空間に存在するWebコンテンツを自動的に収集する機能と，収集したテキストをインデックス化する機能を利用して，Webコンテンツを網羅的に検索するシステムを構築できる．今日インターネット上で著名なYahoo，Goo，InfoSeekなどの検索サイトは，こうした検索エンジンを利用している．

専用検索エンジンの特徴は，コンテンツを収集するサイトを専門家が選別しているので，収集される情報の信頼性が一様に高いことである．こうして作成された専用検索エンジンは，専門分野に特徴的な情報を検索する際に，威力を発する．実際に，我々の研究部門では，医薬品や一般化学物質の安全性情報を収集する目的で，日常的に専門エンジンを用いており，業務の効率化に大きく貢献している．

[*1] （株）アドイン研究所
[*2] （株）インターネットアトラス
[#] To whom correspondence should be addressed:
Tsuguchika Kaminuma; Kamiyoga 1-18-1, Setagaya, Tokyo 158-8501, Japan; Tel: 03-3700-9540, Fax: 03-3700-7592; e-mail: kaminuma@nihs.go.jp

我々の場合，検索エンジンプログラムは，OpenText社のOpenTextを用いた．インターネット上のコンテンツ収集ロボットプログラムは，フリーソフトウェアのWget（GNU Archive）を用いた．マシンはSun Workstation UltraSPARC（Solaris 5.7, memory 512MB, HDD 100GB）を用いた．

つぎに，一般の検索エンジンの仕組みについて多少説明しておく．全文検索は，テキストの内容を構造化することなしに，テキスト全文に対して，そこに含まれるキーワードやフレーズが検索できる仕組みである．全文検索エンジンは，対象となるテキストに対して，その中に含まれる全部の単語をあらかじめ選び出し，単語をキーとしてテキストをインデックス化する．このインデックスによって全文の検索が可能となる．インデックスの作成方法はエンジンの種類によって異なる．OpenTextの特徴は，単語を取り出さずに，テキスト中の全バイトストリームをそのままインデックス化することにある．この方式は，マルチバイト言語を含めてあらゆる言語での検索が可能であることと，形態素解析により不十分な単語区切りがあった場合の，インデックス漏れの心配が無いという利点を持つ．逆に欠点は，インデックスが大きくなり，インデックス化の処理時間も長いことである．Wgetはコンテンツ収集ロボットとして広く用いられているプログラムで，収集する階層の指定，ファイルの種類の指定，階層の深さの指定など，きめ細かい設定が可能である．専門分野に特化した検索エンジンを作成する際には，このような細かい収集オプションの指定ができる機能が重要である．

本システムは，オペレーティングシステムのスケジューリング機能に組み込まれており，定期的なコンテンツ収集とインデックス更新が自動的に行われるように設定されている．

結　果

国立医薬品食品衛生研究所のWebコンテンツのための検索エンジンを開発した．この検索エンジンは当研究所で発信している全Webコンテンツを対象としている．本検索エンジンは当研究所のホームページ（http://www.nihs.go.jp/index.html またはhttp://www.nihs.go.jp/index-j.html）にリンクされており，国立医薬品食品衛生研究所が発信する情報の，道案内の役割を果たしている．インデックスは毎週，最新の情報に自動的に更新されている．

さらに専門エンジンとして，内分泌かく乱物質関連情報用，医薬品関連情報全般用，医薬品緊急安全性情報用，GINC（Global Information Networks on Chemicals）用，健康危機管理関連情報用の検索エンジンを開発した．これらの開発においては，化学物質情報部の研究員がそれぞれ専門の分野について，信頼できる情報発信源を調査し選別した．それぞれの検索エンジンでコンテンツ収集の対象としているサイト数を，他の特徴とともに表1にまとめてある．

Table 1. Domain specific search engines developed in this research.

専用検索エンジンの対象	収集サイト数	インデックスサイズ	更新間隔
NICIのWebコンテンツ	17	8M byte	毎週
内分泌かく乱物質関連情報	18	1.0G byte	毎月
医薬品関連情報全般	82	700M byte	毎月
医薬品緊急安全性情報	42	1.1G byte	隔日
GINC (Global Information Networks on Chemicals)	170	2.0G byte	隔週
健康危機管理関連情報	122	1.7G byte	隔週

信頼できる情報発信源として選別されたサイトは，(1) 公的機関およびそれに準ずる機関，(2) 雑誌，(3) 民間機関で評価データを提供しているサイト，であった．選別の判定基準として，(1) 定評のある印刷物として出版されている内容がWWW上にも掲載されている場合は採用する，(2) 総説を多く掲載している雑誌は優先して採用する，(3) 5-6年連続して調査している間に，途中で消滅することなしに，継続的に情報が更新されているサイトを優先して採用する，(4) 他の機関によって，優れたサイトであると評価された場合は，その結果を十分参考にする，といった事項を考慮した．

表2，3には，我々が開発した専門エンジンが，いかに専門分野の情報の検索に有効であるかを，商用の検索エンジンの検索結果と比較して示す．例として検索文字列「バイアグラ」と「副作用」を用い，バイアグラの副作用情報を検索することを試みた．比較した検索エンジンは，我々が開発した医薬品関連情報用検索エンジンと，商用の検索エンジンであるYahoo Japanである．表2はYahoo Japanの検索結果である．Yahoo Japanの場合，ヒット率1位は医薬品の安全性情報であるが，それに続く上位2位から5位まではすべて，バイアグラの輸入仲介業者が提供している情報であった．2位から5位までは，バイアグラの副作用情

Table 2. Comparison between domain specific and commercial search engines. Retrievals from Yahoo Japan (http://www.yahoo.co.jp/).

順位	タイトル	情報発信源	内容例
1	バイアグラ医薬情報	ファイザー製薬	この度承認された勃起不全治療薬バイアグラに関する情報を公開する公式サイトです。
2	バイアグラを電話かFAXで注文	輸入仲介業者	バイアグラをアメリカより格安に輸入代行。約1週間でお届けできます。
3	バイアグラ薬	輸入仲介業者	こちらでは医師がオンライン診断後処方しますから安心して飲めるのです。
4	漢方バイアグラ	輸入仲介業者	中国漢方界、総力を挙げて開発した、漢方バイアグラを輸入出来るようになりました。
5	マカとバイアグラの比較表	輸入仲介業者	マカとバイアグラの比較表、効果発現速度、効果持続期間

報を調べる目的には全く役立たないサイトである．この例は，商用の検索エンジンを用いると，検索結果に無駄が多く，そのために，検索結果から有用な情報を選別する労力が増えることを示している．こういった傾向は，社会的に話題性の高い医薬品や一般化学物質に，特徴的である．表3は，我々が開発した医薬品関連情報用検索エンジンの検索結果である．ヒット率上位1位から5位まですべてが，バイアグラの副作用情報であり，検索の目的と一致している．

Table 3. Comparison between domain specific and commercial search engines. Retrievals from the drug information specific search engine (http://search.nihs.co.jp:8010/drug-search/).

順位	タイトル	情報発信源	内容例
1	No title	UMIN（大学病院医療情報ネット）	クエン酸シルデナフィル(バイアグラ)と硝酸薬の併用による重篤な副作用について。
2	医薬品等安全性情報155号	UMIN（大学病院医療情報ネット）	バイアグラ錠 25mg(ファイザー)他[慎重投与] カルペリチドを投与中の患者
3	医薬品等安全性情報149号(概要)	UMIN（大学病院医療情報ネット）	米国等において「バイアグラ」の商品名で勃起障害治療薬として発売されている。米国に於いて本剤は硝酸薬との併用が禁忌になっているが
4	topics	国衛研	FDA 認可薬バイアグラ情報、バイアグラに関するドクターレター，日本語訳(一部)(日本薬剤師会)
5	99/08/30 医薬品等安全性情報156号(概要)	UMIN（大学病院医療情報ネット）	クエン酸シルデナフィルは、勃起不全治療薬として本年1月に承認され、これまでに心筋梗塞10例を含む33例の報告があった。

専門家が選別した信頼できるサイトの例として，内分泌かく乱物質関連情報用検索エンジンの対象となっているサイト群の一部を表4に示してある．国際機関や国，地方自治体の研究機関が選別されている．

Table 4. Examples of information sites selected by specialists, in the case of the endocrine disruptor information specific search engine.

厚生省
環境庁
国立環境研究所
東京都環境衛生局
東京都立衛生研究所
東京都環境科学研究所
神奈川県環境科学センター
OECD
IPCS/WHO
EPA
NAS (National Academy of Sciences)
英国環境庁
英国農水省
WWF Canada など

考　察

我々の部が最初に注目した検索エンジンはHarvestであり，つぎにしらべたのがGlimpseであった．Harvestは Glimpseより高機能であったが，日本語に対応していなかった．一方，Glimpseには日本語対応機能があったので，最初自作したNICIのウェブコンテンツの検索エンジンには，Glimpseを用いた．ハードウェアはSun SPARC station20であった．しかし，このシステムをGINC関連サイト専用の検索エンジンに適用してみると，応答速度があまりに遅いことが判明した．今回開発したシステムは，これとは桁違いに強力であり，十分な実用性をもっていることが実証された．

検索エンジンを開発するには，

(1)　強力なハードウェアと高機能のソフトウェアが揃えられること，

(2)　対象となる情報領域に関し，適切なWebサイトを知っており，しかもそれを常時更新していけるその分野の専門家がいること，

(3)　検索エンジンを実際に使っているユーザの声を反映して，システムを改良していける体制がとれていること

が条件である．

我々の研究部には，領域ごとのWebマスターを兼任している専門家が揃っていたことが，専用検索エンジンの開発に幸いした．このようにわれわれの開発基盤は十分に強力なので，他の分野に特化した同様な専用検索エンジンを，情報部の協力のもとで，他の部門が開発することも可能である．

表2，3の例が示すように，今回開発した4種の専門エンジンは，その専門分野の情報を検索するのに，たいへん有用であることが分かった．こういった専用検索エンジンには，2種類の有用な利用法があると考えられる．ひとつは，専門分野の情報を信頼して検索できるシステムとして，一般に公開する利用法である．これは医薬品や一般化学物質の安全性に特定した，情報のポータルサイトとしての利用法である．今日，医薬品や一般化学物質の安全性問題について社会全体の関心が高まっているが，「どこに信頼できる情報があるのか分からない」という疑問に多く接する．この意味で，医薬品や一般化学物質の安全性問題に関する専門家が，信頼できる情報源を選別し，それらの情報について検索できるシステムを提供することは，国研の必要な業務であると位置付けられよう．

もうひとつの有用性は，専門家が，インターネット上の情報を検索する際に発揮される．今日，インターネットは情報の収集源として欠かせないものとなっている．最新の情報を時間差なく獲得できるという意味では，むしろ紙媒体の情報源より有用性が高い場合が多い．実際我々の研究部でも日常的に，インターネットを情報の収集源として利

用している．検索の際に，頻繁に参照する複数サイトを同時にクロス検索することができれば，たいへん効率が良い．これは，頻繁に参照するサイトをまとめて，検索エンジンを作成することにより実現できる．我々の研究部では，今回作成した検索エンジンを，部内でこの目的にも利用しており，日常の業務の効率化に役立てている．

将来課題

インターネットではある関心領域への玄関口となるサイトをポータル（サイト）と呼んでいる．我々の部は，国立衛研がCEOであろうとしている分野におけるわが国のベストポータルの構築をめざしている．関心領域ごとの専用検索エンジンは，その重要な要素技術である．インターネットの世界ではポータルの重要性，さらに検索エンジンの重要性が次第に認識され始めており，技術も激しく進歩している．

我々が課題の一つとして考えているのは，検索の枠に入力する文字列の柔軟性である．例えば，英語と日本語との対応，類似語や上位，下位概念による検索，文字列の柔軟なマッチングなどである．もう一つの改良点は，情報コンテンツをカテゴリーごとに整理して配列することである．現在，開発したシステムを利用してもらいながら，これらの改良を検討している．

謝　辞

東京都臨床研（現，東京都立衛生研究所）の灘岡陽子氏にはHarvestとGlimpseの機能を分析し，検索エンジンへの適用を検討して頂き，数々の有用な助言を頂いた．また石川恵司氏（現石川電機）には，最初の検索エンジンを試作して頂き，今回のシステム開発においても，その初期段階で協力して頂いた．国立医薬品食品衛生研究所の山本都氏と山本美智子氏には，専門エンジンのための情報収集サイトを選別して頂いた．日商岩井インフォコム（株）の安井剛氏には，OpenTextの利用について，技術的なアドバイスを頂いた．また化学物質情報部第２室の中野達也氏，小峰啓氏，中田琴子氏には，システムの環境整備で協力して頂いた．ここに感謝する．

英文要旨和訳

電算機空間は誰も想像すらできなかったような速度で広がり続けており，その空間内を効率良く効果的に探索する技術が重要となっている．検索エンジンは，この問題のひとつの解決方法である．すでに多くの商用検索エンジンが利用可能となっているが，専門家による専門分野の利用においては，重装備であり過ぎて，扱いにくい．そこで我々は，いくつかの専門分野に特定した検索エンジンの開発を試みた．マシンは専用ものを用意し，プログラムは市販のソフトウェアであるOpenTextを用いた．本研究では，NICIのWebコンテンツ用，内分泌かく乱物質関連情報用，医薬品関連情報全般用，医薬品緊急安全性情報用，GINC（Global Information Networks on Chemicals）用，健康危機管理関連情報用の専用検索エンジンを開発した．これらの検索エンジンは，すべてNICIのWebページから，試用に呈している．