学術情報流通における深層ウェブの実態
― 機関リポジトリに収録された文献を用いた調査 ―

安形輝(亜細亜大学)*　　宮田洋輔(慶應義塾大学)
池内淳(筑波大学)　　上田修一(慶應義塾大学)
*agata@asia-u.ac.jp

1. 深層ウェブとその実態調査
1.1 深層ウェブとは

深層ウェブとは、端的に言えば、検索エンジンからはアクセスできないウェブページを指す。検索可能な表層ウェブの対義語である。

「深層ウェブ」という語の提唱者であるBergman[1)]によれば、他のページからリンクでは到達不可能なページ、データベースのようにパラメータ付きクエリーに対し動的に生成されるページ、スクリプトやFlashによって動的に生成されるリンクからしか到達できないページなどとされていた。しかし、検索エンジンのクローラが動的に生成されるページも収集するようになってきた。一方で、ウェブの規模が増大するにつれ、公開されていてもクローラによる収集がされず、検索できない静的なページも増加していると思われる。

本研究での「深層ウェブ」は、動的生成ページなどの観点は排除し、単に検索エンジンで検索できないウェブコンテンツとして扱う。

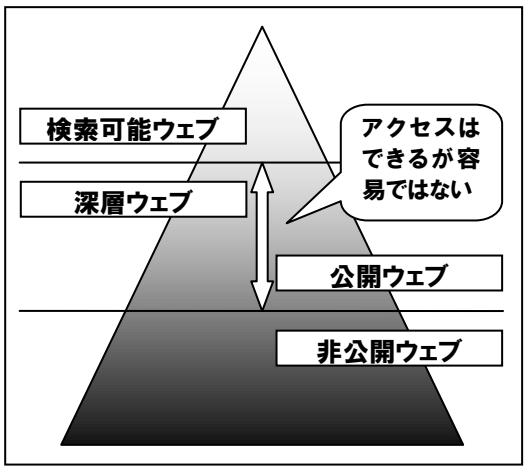

図1 検索可能ウェブと深層ウェブ

ウェブ全体は図1のように「公開されているか」から公開ウェブと非公開ウェブに分けられ、公開ウェブは「検索できるか」という点から、さらに検索可能なウェブと深層ウェブに分割される。

1.2 深層ウェブの重要性

ウェブ上の情報利用行動の多くが検索エンジンから始まる現在、検索エンジンから検索できないコンテンツ(深層ウェブ)は、アクセスはできるが、そのアクセスは容易ではなく、多くの利用者にとって無きに等しい。また、研究者の情報利用行動においても従来のデータベース利用から検索エンジンへの比重が高まりつつあり[2)]、研究業績が深層ウェブに埋没しているかは研究者にとって非常に重要な問題といえる。

1.3 機関リポジトリによる深層ウェブ調査

深層ウェブの実態を把握することは一般的な利用者、研究者にとって重要であるが、その調査はきわめて困難である。なぜなら、ウェブコンテンツの収集の最も一般的な手法はクローリングであるが、大規模な検索エンジンのクローラが収集できないウェブコンテンツを、同じ手段で網羅的に収集することは現実的に不可能なためである。

しかし、近年、機関リポジトリからOMI-PMH(The Open Archives Initiative Protocol for Metadata Harvesting)[3)]によるハーベスティングによってコンテンツの網羅的収集を行い、どの程度、検索可能かを調べた深層ウェブ調査が二つ実施された。McCownら[4)]は2005年6月にOAIsterを網羅的にハーベスティングし、メタデータ中のコンテンツURLから無作為に1,000件を抽出し、Google、Yahoo、MSNで検索した結果、深層ウェブの割合は21%とした。また、Hagedornら[5)]は、McCownらの補足的な調査として2008年6月にGoogle Research APIを使うことで、大規模に調査を行いGoogleのカバー率は44.35%であることを明らかにした。

しかし、これらの調査には不十分な点がある。まず、McCownらの調査は検索エンジンを複

数対象とし、各エンジンの重複を分析しているが、調査URL数が非常に少ない。Hagedornらの調査では調査 URL 数は多いが検索エンジンは一つであるため、重複を分析できない。また、いずれの調査においてもメタデータの収集時期と調査時期のインターバルが空いていないか、明らかではない。そのため、ある検索できないコンテンツがあったときに、クローラの巡回のインターバル中に収録された新規コンテンツだったのか、深層ウェブに埋没しているものなのかの判断ができない。

1.4 本研究の目的

本研究は、先行研究の不十分な点を改善した形で、日本の機関リポジトリにおける全文コンテンツを、Google などの主要な検索エンジンでどの程度検索できるかを調査することで、検索エンジンのカバー率を算出し、深層ウェブの実態を明らかにすることを目的とする。

さらに、機関リポジトリを公開されたウェブのサンプルと仮定することで、検索可能なウェブ／深層ウェブ／公開ウェブの規模について大まかな推計を試みた。

2. 調査環境

2.1 機関リポジトリのハーベスティング

今回の調査対象とするウェブコンテンツは junii2 メタデータフォーマット[6]に対応した日本の 92 の機関リポジトリで公開されている全ての全文ファイルの URL とした。

機関リポジトリの全文ファイルを対象とした理由は、1）OAI-PMH によるハーベスティングにより、リンクを辿るクローラとは別の手法により、全てのコンテンツを収集できること、2）研究成果の公表を主たる目的とするためクローラ排除プロトコルは比較的緩く設定してあること、3）学術情報の全文ファイルは機関リポジトリの中心的なコンテンツであること、である。

日本の機関リポジトリに限定した理由は、junii2 メタデータフォーマットに対応するデータ収集を行うためである。このフォーマットは他のメタデータフォーマットと比較し、情報量が多く、収集後の詳細な分析が可能となる。

機関リポジトリからの網羅的なメタデータのハーベスティングは 2009 年 4 月 11 日に実施した。メタデータから全文ファイルの URL を抽出し、合計で 404,431 件が得られた。

2.2 検索エンジンを対象とした調査

2.2.1 対象とした検索エンジン

調査対象とした検索エンジンは、Google、Yahoo! Japan(「Japan」であるが、登録されているウェブコンテンツは日本のものだけではない)、Bing とした。選定のさいには、1）検索エンジンシェアの上位であること[7]、2）国内だけでなく世界的なサービスを展開していること、3）検索エンジンをプログラムから利用するための API を公開していること、4）任意の URL からの検索を実行可能であること、を条件とした。

2.2.2 検索エンジン API とその利用法

調査に用いた検索 API は Google は Google AJAX Search API、Yahoo!は Yahoo!デベロッパーネットワークのウェブ検索 Web API、Bing は Bing API 2.0 である。各検索エンジン API には検索式として調査対象 URL を渡し、検索結果が 0 件であれば検索できない URL とし、1 件以上であれば検索可能 URL とした。

なお、Yahoo!は全角文字の一部が含まれる URL (10 件)について、検索 API 経由の検索、手動での検索の両方において適切な検索ができなかった（「PDF」という検索式に自動的に変換されてしまう）。それらの URL は検索されなかったものと見なした。

2.2.3 調査実施時期

2009 年 9 月 6 日から 8 日にかけて、404,431 件の全文ファイルの URL を検索式として、検索エンジンに対する調査を実施した。

メタデータの収集と検索エンジンに対する調査の間に 5ヶ月という期間を空けた理由は、先行研究では考慮されていないクローラの巡回頻度の影響を最小限にするためである。つまり、一定期間を空けることで、クローラ巡回の間に新規に追加され、検索できない状況を避け、一定期間以上公開されているが検索できないものを調査可能にするためである。

2.3 カバー率と重複

検索エンジンで検索可能なウェブの割合を示すカバー率と、検索エンジン同士の重複率を算出することで、深層ウェブがどの程度あるかを明らかにすることができる。

カバー率は以下のような式で算出した。

$$\text{カバー率} = \frac{\text{検索可能 URL 数}}{\text{調査対象 URL 数}}$$

また、複数の検索エンジンで検索可能な部分を重複として、重複率も算出した。

3. 深層ウェブの調査結果

3.1 検索エンジンのカバー率

表1は、検索エンジンごとに、調査対象 URL のうちの検索可能 URL 数、さらにその値からカバー率を算出したものである。

表1 検索エンジンのカバー率

	Google	Yahoo	Bing	GYB合算
検索可能URL数	215,259	174,805	115,679	291,024
カバー率	53.2%	43.2%	28.6%	72.0%
			調査対象URL数:	404,431

検索エンジンのなかで単独でのカバー率が最も高いのはGoogleであるが、Googleだけでは調査対象 URL の 53.2%の範囲しかカバーできていないことがわかる。Google、Yahoo!、Bing を組み合わせた場合には 72.0%の範囲までカバーできることがわかる。逆に、全体の28.0%は主要な3つの検索エンジンを組み合わせたとしても検索できず、深層ウェブと言える部分となっている。

3.2 検索エンジン同士の重複率

表2は検索エンジンの重複率を算出したものである。

表2 検索エンジンの重複率

	Google	Yahoo!	Bing
Google		54.2%	37.9%
Yahoo!	66.7%		39.1%
Bing	70.5%	59.2%	

この表は、横軸の検索エンジンのカバー範囲が、縦軸のエンジンのカバー範囲と重複する割合を百分率で示している。例えば、Google の 54.2%は Yahoo!と重複し、Yahoo!の 66.7%は Google と重複している。

また、表1，2の値を用いて、公開ウェブに対する検索エンジンのカバー率の大きさを円の大きさで、エンジン同士の重複率の大きさを円の重なり具合で示したベン図が図2である。

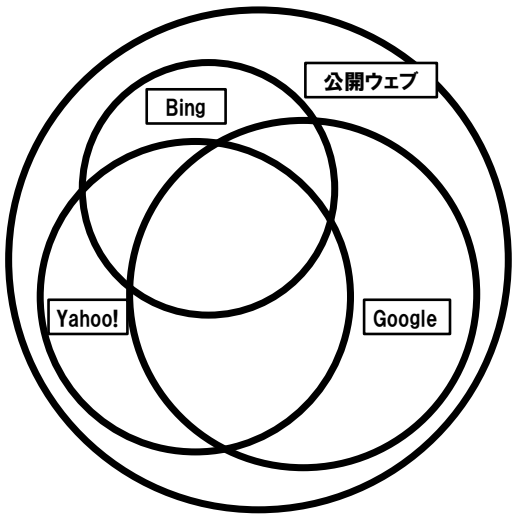

図2 検索エンジンのカバー率と重複率

この図からは、一つの支配的な検索エンジンだけで、ウェブ全体を検索できるわけではないこと、各検索エンジンがカバーする範囲は重複しつつも異なる範囲であることがわかる。

4. ウェブ規模の推計

4.1 ウェブ推計の既往研究

ウェブの規模を推計するためには、Lawrence らによる IP アドレスを用いる手法[8]と検索エンジンを用いる手法[9]が有名である。

前者の手法は無作為な IP アドレスがウェブサーバかを調査し、そのサーバ内にどの程度のウェブページが存在するのかを調査するものである。この手法を用いた比較的最近の 2006 年の加藤らの調査[10]では、動的なページを含め、350 億ページという推計になっている。IP アドレスを用いた手法

表3 検索エンジンの登録数とカバー率に基づくウェブ全体の推計

		Google	Yahoo!	Bing
キーワード	HTTP	4,710,000,000	11,900,000,000	14,700,000,000
	HTML	8,780,000,000	5,090,000,000	5,620,000,000
	A	15,870,000,000	31,100,000,000	6,670,000,000
カバー率		53.23%	43.22%	28.60%
公開ウェブの推計	HTTP	8,849,200,312	27,531,986,499	51,393,387,737
	HTML	16,495,961,516	11,776,286,662	19,648,356,400
	A	29,816,732,262	71,953,342,868	23,319,312,667

の問題点は、あるIPアドレスがウェブサーバだったとしてもトップページから全てのコンテンツをクローリングできるわけではないことである。

検索エンジンを用いた手法は、検索エンジンのカバー率と重複率を利用して、検索可能なウェブの規模を推計するものであり、2000年前後は比較的活発な領域であった。しかし、この手法には複数の検索エンジンのクローラの独立性を仮定している点、検索に用いられる登録数が少ないキーワード集合の適切性など、検索エンジンを用いることによる問題がしばしば指摘されてきた。

4.2 機関リポジトリによる推計手法

機関リポジトリの全文ファイルの URL 集合がどの程度、ウェブ全体のサンプルとして適切かを判断することは難しいが、少なくとも一定以上の期間、公開されているウェブコンテンツであると言える。

今回の深層ウェブの調査結果では、検索エンジンごとにカバー率を算出しているため、検索エンジンを用いたウェブ規模の推計手法を適用することが可能である。また、カバー率の算出自体には検索エンジンを用いていないため、先行研究で指摘されてきた検索エンジン由来の問題が解消されている。

機関リポジトリをウェブ全体のサンプルと仮定し、ウェブ規模の推計を試みた。

4.2 ウェブ規模の推計結果

以前とは異なり、現在、検索エンジンはその登録数(インデックスの大きさ)を公開していないが、検索エンジンを用いたウェブ規模推計のためには、登録数が必要である。そこで、どのウェブページにも含まれると思われるキーワードを用いて登録数の規模を調査した(2009年9月7日)。検索エンジンごとに登録数が多いキーワードは異なるため、複数のキーワードを用いた。キーワードによる登録数を表3の上部に示した。

各検索エンジンのキーワードごとの登録数とカバー率を用いて、深層ウェブも含めた公開ウェブ中の全数を推計したものを表3の下部に示した。

公開ウェブの推計がもっとも大きくなるのは、Yahoo!において、キーワード「A」の登録数とカバー率を用いた場合で、約719億というウェブコンテンツが存在することになる。また、公開ウェブの規模と主要な検索エンジン3つを組み合わせたカバー率(72.0%)を用いると検索可能なウェブは約 518 億、深層ウェブは約 202 億となる。

5. まとめ

本研究では、機関リポジトリに含まれる全文URL 集合を用いて、深層ウェブの実態調査を行った。カバー率は、最も高い Google で 5 割、また、主要な検索エンジンを組み合わせた場合には、カバー率を約 7 割まで上げることでき、これを検索可能なウェブと考えるならば、深層ウェブは約 3 割といえる。

また、機関リポジトリに関する検索エンジンのカバー率と登録数から、ウェブ全体の大まかな推計を行い、最大で 730 億のコンテンツが存在するという結果が得られた。

【注・引用文献】

1) Bergman, M.K. “The deep Web: Surfacing Hidden Value”. The Journal of Electronic Publishing, vol.7, no.1, 2001 (http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104)

2) 例えば、Research Information Network. Researchers and discovery services(http://www.rin.ac.uk/researchers-discovery-services)では研究者がよく利用する情報源はGoogle となっている。

3) Open Archives Initiative Protocol for Metadata Harvesting (http://www.openarchives.org/pmh/)

4) McCown, Frank; Liu, Xiaoming; Nelson, Michael L.; Zubair, Mohammad. "Search Engine Coverage of the OAI-PMH Corpus," IEEE Internet Computing, vol. 10, no. 2, pp. 66-73, Mar./Apr. 2006, doi:10.1109/MIC.2006.41

5) Hagedorn, Kat; Santelli, J. "Google still not indexing hidden web URLs" D-Lib Magazine, vol.14, no.7/8, 2008 (http://www.dlib.org/dlib/july08/hagedorn/07hagedorn.html)

6) メタデータ・フォーマット junii2 (http://www.nii.ac.jp/irp/archive/system/junii2.html)

7) “Top Search Engine Share Trend” http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=5

8) Lawrence, S.;Giles, C.L. “Accessiblity of information on the web”. Nature, vol.400, 1999, p.107-109.

9) Lawrence, S.;Giles, C.L. “Searching the World Wide Web”. Science, vol.280, no.5360, 1998, p.98-99.

10) 加藤真; 山名早人. "Fact of the Web:30 億ページのウェブの解析", DEWS2006, 3B-i6 (2006).