インターネット上の国文学デジタルアーカイブの現状調査とその情報提供

――インターネット検索エンジンを利用して――

北　村　啓　子

要　旨　国文学研究の分野でも、インターネットでの情報発信が急激に増加している。特に電子化された翻刻テキストや原本の影像データなどデジタルアーカイブは研究に利用できる価値ある情報を容易に入手、閲覧することのできる有力な手段である。しかし、インターネット上の膨大な情報の中から必用な情報を見つけ出すことは容易ではないことは誰しも経験しているであろう。インターネット検索システムの現状を報告し、国文学のデジタルアーカイブの所在情報を提供する手段を考案し、その実験の報告をする。また国文学研究者の利用に供するまでの今後の計画を記しておく。

1．はじめに

インターネット上の膨大な情報の中から必用な情報を見つけ出すための仕掛けがインターネット検索システムである。さまざまな手法や機能を備えた検索システムがあり、目を見張る速さで進化している。

国文学分野でも、電子化された翻刻テキストがインターネット上で流通するようになり、また画像デジタル化技術の普及や電子図書館熱も手伝って、大学図書館などを中心に貴重書類の影像データをインターネットで閲覧できるようになってきた。しかしながらこれらデジタルアーカイブがインターネット上のどこにあるかを知る有効な手段はあるだろうか？インターネット検索システムで迅速に的確に捜せるであろうか？

組織だって活動しているところは知名度が高く（老舗の情報処理学文学研究会（JALLC）や日本データベース研究会など)、個人でパワフルに活動してる研究者は同じ専門家の中では人的ネットワーク経由で充分知られているであろう。

しかし個人でも情報発信が容易にできるのがインターネットの長所の一つであり、特に国文学の専門的な研究活動は個人単位で行われる。大学など組織のホームページを利用できれば比較的認知されやすいが、国文学研究者はインターネットサービスプロバイダに個人ホームページを持っている場合も多い。知名度の高い登録サイトを利用して積極的に宣伝すれば別だが、一般にはリンク集に載るなど周知されるまでは期間がかかるであろう。

そこで国文学デジタルアーカイブについては、インターネットに載せると自

動的に迅速にその所在情報を収集し利用者に提供できる仕掛けの開発に取り組むことにする。

本稿ではまず、必須の道具になってきたインターネット検索システムの入門的解説をする。既に使いこなしている人は省略してください。

2．インターネット検索システムについて

インターネット上の欲しい情報があるサイトを捜す検索システムには大きく分けて２つの手法がある。

a．ロボット型ページ検索

ウェブロボット、スパイダーなどと呼ばれるインターネット上のホームページを見つけて情報を収集するプログラムを定期的に走らせて、集めた情報から全文検索エンジンを使って自動的に索引を作成する。利用者はその索引を使うことによってインターネット上の全ホームページ上を捜すのと（ほぼ）同じことができる。母集団が大きくかつ索引の更新が頻繁で新しいホームページを早く周知することができる強力な方法である。しかしホームページ全体から自動的に索引を作る技術はまだ発展途上であり、人手で作った紹介の要約文を使用するのに比べると索引の精度は低くノイズが多い。収集ページの数、その中の抽出採用率によって各検索システムの間に差がある。また索引の切り出し方と構造、使用するページ内の個所（タグ）などの方針に依存する部分が大きく、それが各検索システムの特徴にも、検索時の短所にもなっている。

代表的なページ検索システムで「源氏物語」と「かつ翻刻、電子テキスト、

電子図書館」などの検索結果を比較してみる。（各システムの検索式の書き方に準ずる表記で記述する。検索式で日本語ページ明示的に指定していない検索システムも日本語ページのみを検索している。）

google	源氏物語	25.800件	
	源氏物語　翻刻	238件	# かつ
	源氏物語　電子テキスト	181件	# かつ
	源氏物語　電子図書館	264件	# かつ
goo	源氏物語	15.058件	
	源氏物語　翻刻	216件	# かつ
	源氏物語　電子テキスト	146件	# かつ
	源氏物語　電子図書館	201件	# かつ
Infoseek	源氏物語	15.140件	
	源氏物語　翻刻	189件	# かつ
	源氏物語　電子テキスト	114件	# かつ
AltaVista+源氏物語		12.038 pages	
	+源氏物語 ++翻刻	98 pages	# かつ
	+源氏物語 ++電子テキスト	61 pages	# かつ
Lycos	（源氏物語 AND LANG=jpn）	12.094件	# かつ 日本語ページ
Excite	“源氏物語”	316.828件（.jp内）	
	“源氏物語” AND “翻刻”	565件（.jp内）	# かつ
	源氏物語 電子図書館	2290件（.jp内）	# かつ
OCN navi	源氏物語	7.263件	
InfoNavigator	源氏物語 AND LANG=jpn	7.453件	# かつ 日本語ページ

収録数はgoogleが一番多いようでである。Exciteは同じページを複数カウントする場合があるため数が多く出ている。

検索機能もand, or, boolean,正規表現だけでなく、フレーズ検索、ストップワードも当たり前になってきている。同じURLが何度も出て来る例もあったが、今はどこもURL重複は除いている。サイト毎にまとめて表示できるのも当り前になりつつある。

特徴的な機能を紹介しておく。

・サイト毎にまとめて表示（代表ページだけを表示しサイト内結果へリンクしている）　google Infoseek Lycos
・サイト順表示　Lycos InfoNavigator
・キャッシュ機能：ホームページがダウンしていても採取時にキャッシュ（格納）しているページを見られる google
・関連ページ：リンクされているページを検索、人気度順（リンクされている度合）　google InfoNavigator
・フレッシュなページ（FreshEyeの目的)、日付順表示 goo
・マルチメディア：データタイプを指定して検索できる画像（.jpg .gif）音声（.mp3 .wav）ビデオ（.mpg .swf .mov .avi）など　goo Lycos AltaVista
・ワードナビゲータ：関連するキーワードを提示してくれる
　Excite InfoNavigator
・人名辞書、固有名詞辞書　goo
・リンク先などのtag指定（link: site: title: url:）
　InfoNavigator goo Lycos
・キーワード間の優先順位　goo

検索結果の表示される順番は重要な要素で、期待したページが先に出てくれば上位だけ見れば事足りるのであり、気が効いている感を与える。どのシステ

でも「適合率」という考えを導入し、スコアリングしている。どう「適合」しているかの定義は各々異なり公開されないことが多い。概ね検索キーの出現頻度、複数キー間の距離などは使っていると思われる。個別のアイデアは、例えばgoogleでは公式サイトを優先、リンクされている頻度を利用などを導入しているようで、それぞれ凌ぎを削っている。

ではどの検索エンジンが最適だろうか？目的、分野、捜す対象などによって異ってくるであろう。機能だけではなくインタフェースも重要な要因であり、かつ個人の好みにも影響を受ける。後述する「メタ検索」という一つの条件を複数の検索エンジンのシンタックスに変換して渡してくれる便利なツールがあるので、それで比較してみると良いであろう。

検索システムによって異なる点に、URLの変更、削除に伴い索引を更新するものとしないもの（溜め込んでいくだけ）がある。またURL自身を索引に採用する／しないの差もある。利用時に注意されたい。

日本語ページの索引作りは、日本語固有の問題がそのまま反映される運命にある。自然言語処理技術の進展が全文検索システムへ応用される成果は目覚しく、ページ検索の索引作りにも活かされている。辞書を使って形態素解析を行う新しい手法のインターネット検索システムも出てきている。Gnuライセンスのnamazu（高林哲氏）がサイト内検索システムの構築が容易に行えかつ機能も高く、普及している。

b．ディレクトリ検索

ホームページを分野別に分類し、ディレクトリ構造に並べたもので、典型的

草分けがyahooである。まさに電話帳のイエローページであり、分類のノウハウとサイトの多さが評価の分岐になる。登録制と（抽出）選択制があるが、分類作業は人手で行うため、効率の悪さは避けられず、母集団の小ささ、偏りは避けられない。反面、分類と紹介の要約文を人が書くためホームページ自身に書かれている記述に比して均一されていることが評価できる。

代表的なディレクトリ検索で「源氏物語」を検索してみる。

yahoo:

芸術と人文 > 人文 > 文学 > 詩 > 和歌 > 歌人 > 歴史上の歌人 > 紫式部　14件

Infoseek:　・カルチャー&ホビー→文芸 > 古典　20件

・学び→学問・専門分野 > 文学 > 国文学　35件

goo: 教育&学校 > 人文科学 > 文学 > 源氏物語　4件

Lycos: 芸術と人文源氏物語科学 / 文学 / 日本文学 / 上代・中世文学 / 源氏物語　7件

Excite: 教育＆学校＞人文科学＆芸術＞文学＞古典＞源氏物語　32件

OCNnavi: （階層構造でない）

が出でくる。全検索システムが登録制と選択制（厳選、お勧めなど）を採用している。

検索システムの比較については、ディレクトリ検索で捜すと、

yahoo: ホーム > コンピュータとインターネット > インターネット > WWW > ホームページの検索 >サーチエンジンの比較　10件

lycos: ホーム / コンピュータ・インターネット / インターネット / サイトの検索・リンク集 / 検索エンジン /比較・最新動向　2件

があり、新しい情報が入手でき、参考になるので紹介しておく（ディレクトリ

検索が役立つ例でもある）。

c．メタ検索

同じ条件を複数の検索システムの条件文シンタックスに変換して渡してくれる仕掛けである。

yahooでディレクトリ検索すると、

ホーム ＞ コンピュータとインターネット ＞ インターネット ＞ WWW ＞ ホームページの検索 ＞ メタサーチ　23件

出てくるので、覗いてみて気に入ったのを使ってみるといいだろう。

d．分野別大規模リンク集のディレクトリ化

最近特に目を引くのが、最初はパワフルな個人が蓄積していたリンク集だったのが、多くの人が参照するようになり、複数の個人リンク集がお互いに参照されながら巨大化し、やがて分類されて、特定分野のディレクトリ検索の方向に成長していることである。国文学関連で代表的なのは東北大学文学部後藤斉氏、福井大学岡島昭浩氏、甲南女子大学菊池真一氏などのリンク集をあげられる。

また、個人の専門分野だけではなく、もう少し広い範囲で系統立てて情報収集をおこなっている大規模リンク集も出て来ている。例えば、研究、開発、教育に関する総合リンク集のサイエンスビリッジは非常に詳細に分類され、構造が細分化されている。（実際の手法は知らないが恐らく）システム的にサイト捜しをし、ディレクトリ分類と紹介文を作成する手間をかけており、両者の長所を活かしている。アリアドネ（ARIADNE）や ACADEMIC RESOURCE GUIDEデータベース集成にも広い分野を分類している中に日本文学関連がある。

e．総論

インターネット検索システムは第二期に入ってきていると言ってよいであろう。ディレクトリ検索とページ検索の両方を兼ね備えるのは当たり前になってきた。指数関数的に増加するウェブを検索する索引作りは人手では不可能であり、ロボット型のページ検索は必須であるから。

yahooがページ検索のエンジンとしてgooを採用して便利になったが、最近米国yahoo*はgoogleに切り替えた。yahooも追随するかもしれない。googleはBIGLOBEも採用している。他にも特徴があるがとにかく速さと収集サイト数の多さは注目に値する。ページ検索は索引作りの質と速さが命なので、戦国時代はまだ続きそうである。（本稿が発行される頃はもう新しい検索エンジンが他を圧巻しているかもしれない。）

ページ検索では、もちろん収録ページが多い方が有利である。その次は索引の作り方と検索手法によって結果に差が出てくる。これは捜す物や目的によって適不適があり、一概にどれが良いとは言えない。スコアリングの仕方など見せ方は意外に重要な要素で、これも目的や個人の嗜好にも依るであろう。

目的によっては良く分類されたディレクトリ検索の方が有効な場合もある。現在は両者の対象が重なっているのに別サイトのように扱われているのが不便だが、両者を自動的に融合させることは近い将来可能になるであろう。

世界規模で見るとインターネット上で働いているウェブロボットの種類、数ともに非常に多い。これはネット上の負荷を増大しており、また世界中で重複作業をしているとも言える。ホームページが指数関数的に増加しているのを勘

*国名を断らない限り日本サイトを指す

案すると、処理能力も同じペースで増加していかないと相対的に処理時間が延び、全ホームページを一巡する時間が長くなると、索引の更新ペースが遅くなり、実体と索引の時差が問題となってくる。これらを解決すべくネット上の地域を分担して収集する分散型ロボットの試みも行われている。

誰でも簡単に使えるロボットプログラムもあり、本稿でもそれを利用する。汎用的でかつ精度を上げるのは難しくても、特定分野に絞ることによって精度をあげることは可能である。本稿でもこれを目的としたシステム開発を行っている。興味のある読者にはチャレンジを勧める。

3．目的

国文学研究のリソースとして、インターネット上で利用できる情報、特に電子化された翻刻テキストと原本の影像データの所在を捜すための検索システムの開発を目的とする。漏れを少なく迅速かつ人力をかけずに情報を収集するため、ロボット型のネット検索システムの手法を使う。ただし、ロボット型によるネットへの負荷を極力避け、既存のインターネット検索システムとの重複作業の無駄を省き、効率を上げるため、既存のインターネット検索システムを利用して「作品名」から情報収集対象をせばめる。こうして収集したページの記述内容から「作品名」と記述表現の「パターン」を使って、翻刻テキストまたは影像データの存在を判断する。

次に、収集したページの記述内容を分析することにより、使用した底本や比較参照した異本の情報や翻刻者、撮影した原本の所蔵者などの情報を抽出し、電子化データの素性に関する情報を抽出し、デジタルアーカイブのメタ情報として提供する。さらに、このメタ情報を当館で構築している古典籍の所在情報データベースと照合することにより、原本の所在情報とその電子的複製物であ

るデジタルアーカイブのネット上の所在情報も検索できるシステムの実験を行う。

以上の目的を実現するために次の２つのアプローチを考えている。

ａ．既存のインターネット検索システムを使って古典全般の「作品名」が出現するページを捜し、ロボット型プログラムでページを収集する。→辞書を使って索引作成を行う全文検索エンジンを利用して、古典に関するネット検索システムを構築する。

このアプローチでは前に触れた自然言語処理技術の成果が大きく寄与している。しかし古典を対象にする場合、辞書の不足が問題となる。辞書照合を基本にする手法では、辞書の規模がそのまま評価に効く。古典では12世紀以上に渡る固有名詞、死語、造語など辞書の整備は現代語の比ではない。

しかし当館では創立以来の目録データベースをはじめ各種データベースの構築を通して作品名、人名など固有名詞の蓄積があり、さらにシソーラス、電子化辞書の構築などを目標に様々なプロジェクトに取り組んできており約二十万のオーダーの専門用語も蓄積してきている。これらを全文検索システムでの索引作成に利用することは、今後自然言語処理技術の成果を取り込んでいく上で重要な取り組みである。

ｂ．既存のインターネット検索システムを使って古典全般の「作品名」が出現するページを捜し、ロボット型プログラムでページを収集する。→「作品名」と記述表現の「パターン」を使って、翻刻テキスト、影像データを掲載しているかどうかファイルタリングする。(図１参照)

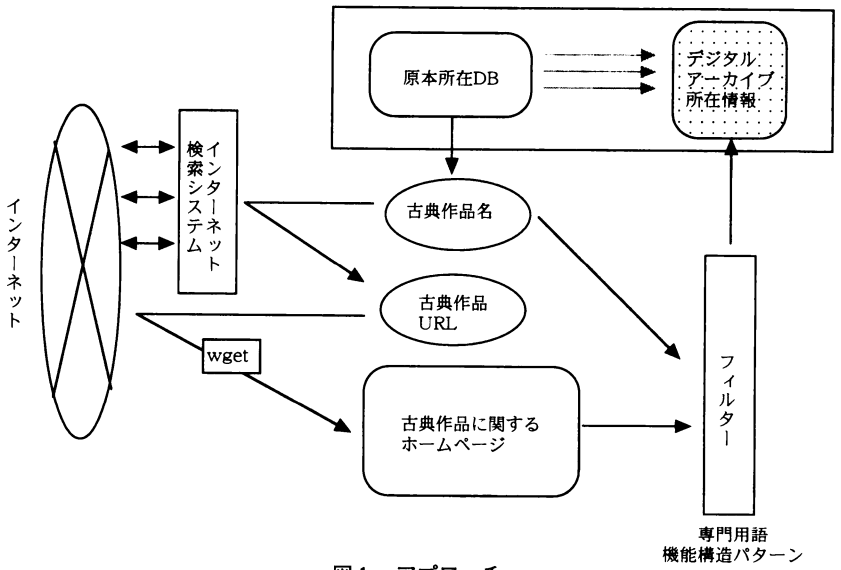

図1　アプローチ

テキスト分析をする際に、機能語に着目した機能構造を利用する手法がある。目的を絞れば手軽なテキスト処理である程度の成果をあげることができ、辞書が必用な重い処理を避けられる。

最近のリンク集には、電子テキストと画像データベース（貴重書コレクション）などに分類されているものが出現してきているので、分類されたページを分析し、機能構造のパターンを抽出してフィルタを構築する。このフィルタを使って、自動的に迅速にデジタルアーカイブの所在情報を収集することができる。

a.のアプローチに似た方法で東京外国語大学永崎研宣氏がHumanities-searchという人文科学系ホームページの全文検索システムを作成運用しているのをインターネット検索システムで発見した。人手で作られたリンク集から人文科学

系ページの所在情報（URL）を抽出し（前半のこの部分が異る)、ロボット型プログラムでそのページを取得し、日本語全文検索システムnamazuを利用している。古語に関する辞書の追加、整備は興味あるところだが未確認である。約5万件の文書が収集、索引化されている。今後収集ページの増加に伴い精度が向上していくことが大いに期待される。

本稿では、b.のアプローチに取り組んだ報告を行う。

4．方法

実験作品として、今年様々な意味で注目を浴びた「源氏物語」を選んだ。インターネット上に沢山のページがあり、翻刻テキスト、影像データを掲載する周知なサイトもいくつもある。占い、パロディ、漫画など新しい鑑賞方法を楽しんでいるページや二千円札の話題など（今回はノイズとなる）も沢山見受けられる。

1. googleのページ検索で「源氏物語」を捜すと25,800件の結果であった。googleは同じURL中の複数ヒットしたページはまとめて、トップページとその中の最初の1ページだけを表示する（該当サイト内の検索結果表示へリンクする体裁を採っている)。結果表示されたのは627件、この中の異なりサイト604件であった。この結果ページからURLを抽出し、wgetを使って該当サイトのページを収集する。wgetはネット上のページを自動収集するロボット型プログラムの一つで、Gnuライセンスで誰でも利用できる。

google 源氏物語 25,800件

源氏物語　校訂　270件 結果表示 204件　異なりサイト 195件

源氏物語　翻刻　238件　結果表示 192件　異なりサイト 215件

源氏物語 電子図書館 264件 結果表示 205件　異なりサイト 196件

2. 国文学関連の著名なリンク集の中から以下に上げるものを選び、翻刻テキスト、原本の影像に分類されているサイトを抜きだし、これらをマージしてから異なりサイトを抽出する。

・東北大学後藤氏「国内人文系研究機関WWWホームページリスト」
・福井大学岡島昭浩氏の「日本文学等テキストファイル」
・甲南女子大学菊池真一氏「日本文学関係テキストファイル等」
・明星大学柴田雅生氏の「電子化された日本語テキスト」
・青空文庫の「日本文学・電子テキストのある場所」
・サイエンスビレッジの「日本古典文学の森」「図書館・博物館とインターネット」
・本とコンピュータ「こんなにたくさんあるぞ 日本文学電子テキスト」(雑誌)
・当館のリンク集、筆者のリンク集

3. 2.のリンク集から抽出した翻刻テキスト、原本の影像データを掲載しているページから、その旨を特徴的に示す説明文の機能構造のパターンとそれを埋めることを想定される語彙を抜きだす。例えば以下のようなものである。
%|...|は目録データベースなどから抽出した想定される一覧リストである。

底本%|書名|に%|所蔵者|蔵を使用した
%|所蔵者|所蔵の%|書名|を撮影した
%|書名|を画像データベース化した
%|書名|（%|所蔵者|蔵）を諸本とし
仮名遣い....異同...音韻....

4. 2.で抽出したページについて、3.のパターン（正規表現を使う）と語彙集

との照合を行う。全ページが抽出できるまで失敗した例を分析して、パターン、語彙集を追加、修正する。

5. 1.で収集したページについて 4.で仕上げたパターン、語彙集との照合を行い、新たに発見されたページについて評価する。

6. 以上のプロセスを経て、構築したパターンと語彙集を使って当館の所在データベースから抽出した約数十万件の作品名について、同じ方法でインターネット上のデジタルアーカイブの所在情報を収集する。

5．結果

今回の実験では、収集ページが莫大になったのでまだ全分析が終了していない。現在までのところリンク集にない発見できたのはわずか3サイトである。

約30種類のパターンを用意し正規表現も使って照合したが、サイトによる表現の差が予想以上に大きく、さらなるパターンの種類収集が必要である。

しかしながら、原本影像データについては、素性情報を発見できなければ有効な語彙は殆んどない。ページの記述（テキスト）から抽出するより、データの種類を使って .jpg, .gifファイルを大量に蓄積しているかどうかを調べる方が単純だが有効である感触を得た。翻刻テキストも、.txtやパソコンのアーカイブ・圧縮形式 .lzhなどでダウンロードできるようにしているサイトも少なからずあった。データの種類と機能構造パターンを使う適当なバランスを見極めることで、認識率、効率ともに上げることができそうである。

原本の解説（所蔵者、書誌的事項）、凡例（句読点、濁点、漢字）、作成プロ

セスの解説、翻刻／訳の担当者などの記述が手がかりになると考えていたが、そういったものが一切ないサイトが結構あった。

検索結果からのリンクはトップの説明や凡例のページを飛ばして直接画像には張られている場合があるのが影響しているようである。サイト内を上向きにも捜す必要がある。

翻刻テキスト、影像データとも、該当ページに掲載されているのか、出版社からオフラインのメディアで販売されている広告なのかの区別が難しい（記述内容が似ている）。サイトのドメイン名や出版社名、価格表示などで外すことも考えられる。インターネット上で利用できるデジタルアーカイブを対象に考えてスタートしたが、データが学術的に利用価値のあるものであれば商用であっても有用な情報であるという考え方もある。いづれにしても区別は難しくはなさそうである。

6．今後

まだ途中結果ではあるが、上で述べた

1．ファイルの種類を使って判断するフィルタリング機能を入れる
2．ドメインによる分類（アカデミック、商用、個人サイトなど）を利用する
3．パターンを追加する

の改良を加え、実験を繰り返す。その結果得られた源氏物語のデジタルアーカイブの所在情報は、使用に耐えるフィルタまで到達した時点で、筆者のホームページから利用できるようにする。引続き、数十万作品の所在調査を行い、これも同ホームページに追加掲載していく。その後の追加差分は少ないので、適度な頻度で更新をしていこうと考えている。

当館の所在情報データベースは原本の所在情報の目録だが、これに翻刻テキスト、原本画像のネット上の所在情報を合わせて検索できる実験システムの開発に取り組む予定である。これも筆者のホームページで案内する。

海外サイトにも、日本語テキストイニシアチブ（バージニア大学&ピッツバーグ大学）のようにアカデミックな活動をしているサイトもある。国内のページ検索システムでも、海外の日本語サイトの索引も作っているシステムも増えて来ているので、それらの利用を検討していく。

調査数が多く分析が充分でない段階で本稿の執筆となってしまった。充分な結果報告ができないことをお詫びする。発行以降、最新情報を筆者ホームページに掲載して行くので、興味のある方は参照ください。

最後に、説明するより実際覗いてみた方が早いことが多いので、敢えて紙面を割いてURLを記入しなかった。発行日に合わせて本稿のHTML版を筆者のホームページに掲載するのでそこで思う存分クリックして頂けることを期待する。

http://www.nijl.ac.jp/~keiko