

インターネット情報の活用

国立国会図書館長
長尾 真

PROFILE

1994年に電子図書館アリアドネを公開。
2000～2007年、日本図書館協会会長。



1

知識の所在の変化

人類の知識は図書館に蓄積されて来た。そしてこれを利用し、新しい知識を創造することが大学や研究所で行われ、その結果が再び図書館に収集されるという循環があった。大学において図書館が重要な位置付けを与えられて来た由縁である。

ところが20世紀の後半に入ってコンピュータが発達し、情報処理能力が高まるとともに、種々の情報・知識がデータベースに貯えられ、これが高速のネットワークによってどこにでも伝えられる時代がくることによって、過去の伝統的な知識生産の構図が変化して来た。

以前は一般的な知識は百科事典から得ていたが、今日ではインターネット上のウィキペディアの情報が非常によく利用されている。ただウィキペディアの内容はそれぞれの項目についての専門家が書いているとは限らないため、しばしば誤った記述が存在するといわれていることに注意すべきであろう。

インターネット上の情報はこのようにならずしも全てが信頼できるものではないが、手軽にいろんな情報を取り出せる知識の宝庫ようになって来ている。したがって今日の知識生産の構図は、インターネット上の知識情報を利用し、社会の“誰も”が新しい知識の生産に参加し、こうして新しく作られた知識・情報がインターネット上に発信されるという形となって来ている。

ただ図書館や各種の専門データベースの知識や情報は、依然として最も確実に信頼の出来るものとして、全ての

知識や情報の中心に位置しているわけで、これらの知識や情報をうまく利用することは大切なことである。

2

ウェブ情報の収集

国立国会図書館は日本で唯一の納本図書館として特異な位置付けにある。すなわち、日本で発行される出版物は我が国の文化財として全て国立国会図書館に納入することを法律で規定している。

ところが近年出版の形態が変わって来て、電子的な世界でのみ出版あるいは発信されるものが増えて来ており、これらは納本制度の枠外にあるので国立国会図書館には入って来ないという状況が生じて来ているのである。しかし、これらの電子世界における情報も日本の文化財の対象とすべきは当然であるので、国立国会図書館はこれを収集する必要があると考えている。

学会が出す電子ジャーナルなどは信頼性が高く安定したものであるが、インターネット上のウェブサイトは多くの場合、個人が自由に情報を書き込み、送り、やりとりしながら書きなおしたりと、非常に不安定なものである。また誤った情報を意図的に発信したり、著作権侵害、人権侵害、名誉毀損、プライバシー侵害、わいせつ物など、問題のある情報も存在する。

国立国会図書館としては、本来的にはウェブ上の情報は何らの選択上の判断もせずに、包括的に自動収集すべきものであるが、そうすることには著作権やその他の問題がある。そこで現時点では国の機関や、都道府県、政

令指定都市、法人・機構、大学等の限られた対象に対して、いちいちそのホームページの情報をコピーして保存し一般の利用に供する許諾をとって集めている。そのサイト数は約1,500であり、これらは集めるべきサイト数からすると微々たる数である。

そこで納本制度と同じように、国立国会図書館が必要とするウェブサイトの情報を自由に収集・蓄積して利用に供することを保証する法律を作る必要がある。現時点では、その必要性はある程度認識されるようになってきているものの、ではどの範囲のものを著作権にかかわらず集めるかという範囲、境界を法律に明示的に表現することが非常に難しく、法律にできずにいるというのが実状である。

3 ウェブ情報の信頼性、信憑性

前節にも書いたようにウェブ情報には信頼できるもの、そうでないもの、宣伝を目的とするものなど、種々の性質のものが混在している。またグーグルの検索をみても分かるように、注意深く検索キーを設定しても取り出されてくる情報は何万、何十万となってしまう、これでは検索した意味がないということになりかねない。

そこで検索結果にさらに高度な言語情報処理を行うことによって、利用者に分かりやすく、安心して利用してもらえる情報提示を考えねばならない。それらは次のようなものである。

- 1) 検索結果で類似の内容のものを一まとめにするクラスタリング操作を導入し、数万の検索結果を数十以下のグループにまとめる。
- 2) それぞれのグループの内容を要約技術を用いて要約し、理解しやすくする。
- 3) その要約された内容がどの程度信頼できるものか、正しい知識に照らしてみた時の信憑性はどうか等を推定して示す。そのためには、情報の発信者の信頼性、学問的知識との整合性等を調べることに

なる。

- 4) 検索された情報のうち、いわゆるロングテールに存在するもので、信頼性の高いもの、あるいは多くの類似情報と対立するマイノリティ情報といったものをすくいあげ、少数派だが重視すべき情報として提示する。

このような内容のものは、現在「情報分析研究」という名称で、京都大学の黒橋禎夫教授、情報通信研究機構などで研究が進められている。このような機能をそなえた情報検索システムはこれからのウェブ検索で必須のものとなるだろう。

4 知識活用の時代

今日知識は図書館だけに存在するわけではない。広くインターネット上に種々の知識・情報が存在する。しかし産業界・学界などに必要な知識は特許情報データベースや各種の学術研究論文データベース、あるいはそれらに関連する数値データベース、画像データベースなどであろう。それらの中でも産学連携が進んで来ている今日の状況をみれば、特許情報データベースを検索し、分析することによって多くのヒントが得られるし、また次の研究開発のトピックなども浮び上ってくるものと考えられる。

人類がこれまでに築いて来た知識をいかに活用するかがこれからのキーポイントである。私は国立国会図書館で“知識は我らを豊かにする”という標語を掲げている。