特集 〈検索〉の未来

検索から連想へ
情報を発想力に変換する連想エンジン

高野明彦（国立情報学研究所連想情報学研究開発センター）
たかの　あきひこ

ウェブ検索は強力で魅力的だが，思考のためのツールとしては問題点も多い．「連想の情報学」では，量よりも質にこだわって蓄積されてきた多様な電子情報を，「連想」で柔らかくつないで自分の記憶を刺激する情報環境を構築している．無意識下で私たちの発想を鍛える新しい思考空間となりうるか．

思考環境としての電子情報空間

インターネットの普及により，ありとあらゆる情報が電子化され，ウェブ・コンテンツはすでに400億ページを超えたともいわれる．私たちが必要とする情報はすべてウェブ上で手に入るとさえ感じられる．主要な学術ジャーナルは電子化され，購読料さえ払えば，ウェブ経由で必要な論文を瞬時に取り寄せて読むことができる．英語圏では，1000万冊以上の書籍を全ページ電子化して，図書館の蔵書全体を丸ごと全文検索可能にしようというプロジェクトも進められている．私たちはこのような大量の電子情報と日常的に向きあい，それらの有効活用なしに充実した知的生活は考えられなくなったといえる．

しかし，ウェブ・コンテンツの内容を具に眺めると，電子ジャーナルや電子書籍は良いとしても，多くは玉石混交である．粗製乱造による質の低下が嘆かれている書籍の世界と比べても，情報の平均的な信頼性は著しく低いといわざるを得ない．どこかに第一級の高信頼な情報が含まれているとしても，大量のゴミのような情報の中からそれらを自力で選別することは難しい．

現在，この困難な仕事に役立つほとんど唯一のツールは，Google や goo などのウェブ検索サービスである．数十億ページを巡回して全文検索可能にするサービスは確かに強力だが，集められた情報はジャンルや文化的背景などを無視した形で扱われる．そこから求める情報を掬い上げる手段は，指定した文字列の有無によるキーワード検索だけであり，限界も多い．大量に見つかる情報のうち，どれをユーザに提示するかは検索エンジン任せで，多くの場合，ある種の「人気順」で返される．当然，「人気」は内容の信頼性を保証するものではない．広がりのあるテーマや漠然とした話題については，キーワードを数語指定する検索では，まったく歯が立たないことをしばしば経験する．自分が直面している問題解決のヒントや自分の思考を深めるのに役立つ示唆を得るのは難しいのが実情だ．

その結果，多くの人は一生かかっても読みきれない量の無秩序な情報を前に呆然として，ただ漫然とネットサーフィンを繰り返すことになる．たまたま求めている情報が得られたとしても，まだどこかにより価値の高い情報が埋もれているのではないかという物足りなさがつねに残り，1冊の本を読み終えた時のような達成感はなかなか得られない．このように，現在の電子情報空間は，思考するための環境としては大変貧弱なものに止まっている．

ウェブ検索の落とし穴

にもかかわらず，私たちはウェブ検索をよく使う．私自身，Google 依存症といってもよいほどである．Google 検索ですぐに見つけられるとわかると，その事柄自体は覚えずに，検索に使うキーワードだけを覚えるようになる．ある課題について考え始めるときに，まず Google をどのようなキーワードで検索すべきかと無意識に考えている．そうやって思いついた数語のキーワードで集めたウェブページが，関連する全知識の要約(縮図)であると錯覚することがある．自分がたまたま思いついたキーワードだという偏りや，検索エンジンが勝手に重要だと判断した規準で選ばれたページ群だという偏りをつい忘れがちである．

Google などのキーワード検索で探した情報源をざっと眺めて，目前の課題解決に役立ちそうな情報をコピー & ペーストすることで，自分の意図に合う情報をいとも簡単に収集できる．危険なのは，この単純な検索作業を，「自分の頭で考える」ことと混同しがちだということだ．Web2.0 という言葉の流行に乗って発信されるウェブページは，そのような「自分の考え」で溢れかえることになる．私たちは Google を自在に使いこなして，自分の自由な意思決定に役立てていると思い込んでいるが，それが本当かどうかはかなり怪しい．

電子情報に限らず，一般に人間が意思決定で陥り易い罠として，次の 5 つがよく知られている[(1)]．(1)アンカリング: 最初に見つけた情報から過度に影響を受ける，(2)確証: 無意識のうちに自分の既成概念を支持するデータを探し，それを覆す証拠は避ける，(3)記銘性: 直近の出来事や劇的な事件には過度に影響を受ける，複数の情報源から繰り返し同じ情報を受け取ると信用してしまう，(4)現状: 現状維持に役立つことを受け入れやすい，(5)埋没費用: 過去の過ちをなかなか認めずに，これまでの選択を正当化する方向で意思決定する．

信頼性の保証がなく，コピー & ペーストで再生産された似非情報が溢れる膨大なウェブページを対象に，Google の強力なキーワード検索機能に頼って情報収集が行われるとき，これら 5 つの罠に陥らないようにするのは至難の業である．私たちは自分の仮説を支持する情報源を効率よく収集して，集まった情報源に繰り返し書かれている内容を真実だと確信する．自分の仮説にあうデータが期待通り収集できるので，たとえ仮説が誤りであってもなかなか気づかない．Google のパワーがこの罠から抜け出すチャンスを封じているともいえる．

さらに悪いことに，最近の「意識」の研究で明らかになったように，人間の本質的な意思決定は，かなりの部分が無意識下に任されている．意識の役割は，無意識下で準備されている可能性の候補の中から，いくつかを拒否することぐらいであるという[(2)]．だとすれば，無意識下で取り込んでいる情報によって，私たちの「自由な」意思決定の範囲が決められているといってもよく，最初に視野へどのような情報が飛び込んできたかによって，私たちの発想は縛られていることになる．このように，電子情報空間における私たちの視線の向け先や視野の広がりは，ビーム幅の狭い強力なサーチライトのようなウェブ検索機能によって強く規定されているのである．

連想の情報学

では，どうしたら電子情報空間をもっと私たちが思考するのに相応しい場所にできるだろうか．私が最も重要だと考える要件は，情報の信頼性と情報への適切な視野の確保である．信頼性については後で検討することにして，まずは広くバランスのよい視野をどうやって得るかを考える．これには，情報を文脈で捉えることが有効である．私たちがある情報を理解するとは，その情報と関連情報との関係性を適切な文脈に位置づけることである．私たちの頭脳は，外から情報を受け取ると，無意識下で記憶を探索して関連情報を収集し，そこから文脈を作り出そうとする．自分の頭の中に，

外に見える情報の文脈と親和性の高い文脈を作り出せたとき，私たちは「わかった！」と感じるのだろう．これが私の考える「思考する」ことの情報処理的な側面である．

人は誕生以来の記憶を恐らくはずっと潜在意識下で持ち続けながら，普段の生活ではそのうちのほんの一部だけを思い出して使っている．つまり，自分の脳内の記憶を連想的に探索し，関連情報を無意識下で想起しながら，知的活動を行っていると考えられている．一方，情報空間には一生かかっても眺め尽くせない大量の電子情報が存在し，そこから私たちの思考に役立つ情報を収集して活用することが求められている．考えてみれば，この電子情報は私たちの潜在記憶に似ていないか．

このような観点から，われわれは「連想の情報学」を提唱して，人が思考するのに適した電子情報空間のあり方を模索してきた．大量の電子情報のプールから，求めている情報と内容的に近そうな情報をざっくりと掬ってきて，その概要を人間の連想を刺激するような形で提示する．それを見た私たちの頭の中では，関連する記憶が無意識に呼び起こされ，それが次なる電子情報空間との対話のきっかけになるのではないかということである．私たちの頭の中に眠る膨大だが意識的にはなかなか活用できない潜在記憶と，確かに存在しているが一度も見たことのない大量の電子情報を，「連想」によって結びつけようという試みである．

最初は情報技術的な裏づけのない思いつきだったが，その後，数千万件規模の文書データベースを自在に扱える連想計算エンジンGETA (**G**eneric **E**ngine for **T**ransposable **A**ssociation)を開発して研究が大きく進展した．

連想計算で「どこからでもリンク」

GETAはコンテンツを受け取り，それに対する連想計算という計算機構を提供する．この意味で，コンテンツというデータ群を連想計算へ変換する一種のコンパイラと考えることができる．連想計算は，情報内容の類似性に基づく関連情報の探索・分析・提示に役立つ．具体的には，任意の文書を受け取って，それと内容的に関連する文書群や単語群を返す計算である．基となっているコンテンツから関連文書を見つけ出す機能を連想検索，それらを要約する単語群の自動抽出機能を関連語抽出と呼んでいる．関連性の評価には，使用されている単語の重複を数値化する内積型の類似性計量であれば，どのようなものでも利用できる．

新聞記事などの自然な文書を丸ごと質問文として与えると，数千万件規模のコンテンツから瞬時に関連文書を連想検索して，関連語を抽出できる．これはちょうど，そのコンテンツに精通している人間に新聞記事を読ませて，その人の思いついた関連文書やその人の頭に浮かんだ言葉が返される気分である．コンテンツ全体はその人の潜在記憶の役割を果たしている．とくに関連語は基となるコンテンツの内容を敏感に反映するので，その話題に関するコンテンツの質を判断する手がかりにもなる．ユーザは返された関連語を見て，忘れていた知識を思い出し，新たな視点を得ることも多い．

ここで重要なのは，連想計算の相互運用性である．連想計算は，外部から与えられた任意の文脈(文書)に対して，コンテンツ内の文書を関連性の強さで順位付けするので，その文脈から関連文書へのリンクが自動生成される仕組みと考えることもできる．したがって，コンテンツをGETAで“コンパイル”すると，外部の情報からそのコンテンツへの関連性リンクをいつでも自動生成可能な「どこからでもリンク」状態になる．これは反応性の高い一種の“ラジカル”状態であり，他のコンテンツといつでも相互に関連付け可能となる．この機能により，私たちは文脈として表現された自分の問題意識を直接多様な情報源にぶつけて，視野を広げていくことができる．

信頼性の基点としての情報サービス

では，高信頼な電子情報を得るにはどうすればよいだろうか．住みやすい街に道路や公園が欠か

せないように，ウェブ上にも日々の暮らしに役立ち，安心して使える公共コンテンツが必要である．少し大げさに言えば，社会全体として「知の公共財」と呼べるような情報サービスがきちんと維持され，広く提供されることがその文化の底力を規定する．最近，学び・思考するための環境がウェブ上に整備され，それを用いた遠隔教育の実践も始まっている．長い年月と多くの労力をかけて維持してきた高信頼なコンテンツをそのような学びの環境で利用可能にすることの意義は大きい．

われわれは，このような学びの入口へ案内するための公共サービスの 1 つを目指して，2004 年 6 月に「新書マップ(http://shinshomap.info/)」を立ち上げた．これはさまざまな話題への入門書として書かれている新書・選書 9000 冊を，トピック別のテーマ書棚に分類したもので，各テーマには本のリストと読書ガイドが付いている．表紙と書棚に並べた背表紙の写真も眺められるので，ウェブ上で書店歩きの楽しさを少し味わえる．個々のテーマには，拾われている本の目次・概要や読書ガイドなど多角的な文書が付与されているので，連想計算に好適なコンテンツになっている．実際，どのような話題の文章を貼りこんで検索しても，10 個の関連テーマが返され，自分が気づかなかった視点を教えられる．

われわれの研究室では，さらにいろいろな情報源を GETA で“コンパイル”し，連想計算の品質について評価している．その中でとくに有望な結果が得られたコンテンツは，図書館のリファレンスコーナーでお馴染みの百科事典や専門辞典である．これらは，各分野で得られている知識を概観するのに役立つ優良コンテンツで，専門家によって書かれているため情報の精度は高い．同じ話題でも，百科事典と専門辞典では，前提としている予備知識や記述の専門性が異なるため，両方を合わせて読むことで多角的な理解が得られることを実感できる．新聞記事データベースや報道写真データベースにも適用して試しているが，ときどき連想の精度は落ちるものの，具体性のある事実の記述に詳しく，主張を裏付ける情報を得るのには大いに役に立つ．

このように，専門家によって編集され，記述内容が精査された情報源のみから情報を収集することは，ある意味ではウェブ普及以前の情報環境に戻るように思われるかもしれない．しかし，先に論じたように，私たちが最初に出会う情報から過度の影響を受けることを考慮すると，電子情報空間を歩き始める最初の一歩は，このように性格づけのはっきりした信頼性の高い情報源から踏み出すのが肝心だと考える．各情報源が，舞台裏のコピー & ペーストで結び付いた似非情報ではなく，独自の判断で収録された独立した情報を提供していることも重要なポイントである．

これらの独立で高信頼な情報源こそが，電子情報空間における信頼性の基点になる．どれだけ豊かで多様な信頼性の基点を利用可能にしているかが，その文化の継承や発展にとって重要だと考えている．

「想・IMAGINE」は情報空間の六分儀になれるか？

専門辞典，百科事典，新聞記事データベースなどの高信頼な情報源の重要性は，一貫した編集方針により長年に渡って情報を蓄えて，取り扱う分野について安定的でバランスのよい見方を提供していることにある．それらは宇宙的な規模で広がる電子情報空間の中で，それぞれ独立の位置を占めて輝く星座のような存在で，その周りの星々を観測する手がかりを与えてくれる．

これらの情報源をそれぞれ GETA でコンパイルしておくことにより，各情報源からの“見え”をいつでも連想計算により求めることができる．複数の情報源からの見えを組み合わせることにより，私たちは電子情報空間の中での自分の位置を知り，どの方向に歩き出すべきか判断できるのである．

このようなコンセプトに基づいて，われわれは「想・IMAGINE」をデザインした．多様で深みのある高信頼な情報源を複数横に並べて，そこに

図1——想—IMAGINE.

ユーザは自分の想いを文章として投げかける．すると，個々の情報源ごとに連想計算されて，その情報源ならではの情報の見え(文脈)が返される．複数の情報源からの見えを書棚のように並べて一覧することで，さらに多角的な文脈(情報の景色)が得られる．ユーザは各情報源の想いを読み解きながら歩き回り，その中から心に響く情報を取り上げて読み進む．その過程でユーザの想いは少しずつ変化し，それに呼応して情報源が示す情報の景色も変わる．想いが連なり連想に変わる．

「想・IMAGINE」技術を用いた情報サービスとして，「想—IMAGINE Book Search」を公開している(http://imagine.bookmap.info/)．現在は本に関わる情報サービス(Webcat Plus，新書マップ，Book Town じんぼう)や文化遺産オンライン，ウィキペディアなど，ウェブ上で無償公開されている7種の情報源を対象に，関連情報を連想検索で一気に収集できる．収集された情報の中から気に入ったものにチェックを入れて，それらを基点に関連情報を再収集できる．

「想・IMAGINE」の真価は，有料・無料の壁を越えて，専門辞典・百科事典・新聞記事データベースなども含めた多様で上質な情報源が利用可能となったときに発揮される(図1)．連想計算が保証する相互運用性により，情報源同士は一切情報交換することなしに，ユーザが複数の情報源を自在に組み合わせて利用できる．Web サービスと AJAX と呼ばれる最新のウェブ技術を活用して構築しているため，各情報源の提供者が世界中に分散していても構わない．ユーザは情報源から返される情報の景色を眺めて判断し，情報源を自由に変更しながら，情報空間を探索できる．視野の広がりや情報の専門性についても，情報源の選択を通じて間接的にコントロールできる．

この「想・IMAGINE」を実社会へ広めるために現実の図書館との連携を進めている．その第一歩として，本年5月に新拠点で改装オープンする千代田区立図書館に協力して，そのリファレンスコーナーに「想・IMAGINE」や新書マップを使った先進的な情報利用環境を構築する．図書館では，百科事典や専門辞典，新聞記事データベースなど「想・IMAGINE」にとって理想的な情報源を無理なく確保できるため，ウェブ上の公開サービスを超えた深みのある情報サービスを提供できる．かつて私たちの思考環境として大きな役割を果たしてきた図書館が，われわれの情報技術で，再び豊かな知的刺激を得るための空間としての魅力を取り戻すと信じている．

知識や真実の価値を信じた先人たちによって作られ，その深い想いが込められた多様な情報源を，今ここに生きる私たちの切実な課題や思考と響き合わせて柔らかく接続する「想・IMAGINE」は，情報空間における新しい六分儀になりうると自負している．さらに技術を洗練させ，さらに多くの文化の記憶を取り込んで，ウェブも含めた情報空間における新しい航海術を確立したい．

文献

（1） P. Morville，浅野紀予訳: アンビエント・ファインダビリティ，オライリー・ジャパン(2006)
（2） ベンジャミン・リベット，下條信輔訳: マインド・タイム，岩波書店(2005)