When Everything Is Searchable

Eric A. Brewer
brewer@eecs.berkeley.edu

すべてがサーチ可能になるとき

翻訳：安藤　進
sando@twics.com

インターネットの焦点がコミュニケーションからサーチに移行した．まず，サーチエンジンのサイトが注目された．サーチエンジンサイトはポータルサイトへ進化してきたが，ネットワークがどうあるべきかやどのように評価すべきかという観点では，サーチそのものが土台にあることには変わりがない．インターネットが技術的にも社会的にもすばらしいのは，何の規制も受けずに自由に情報を共有できることである．誰でも，情報を追加し自分の意見を表明できる．サーチは，普通のユーザでもこのような混沌とした世界に立ち向かい，オンラインで利用可能な数十億のドキュメント，将来は数兆にも達するドキュメントを活用できるための手段のことである．このように膨大な数の情報全体を整理して体系化することは人間にはまず不可能なので，無秩序な状態から必要な情報を探す手段・技術が必要になる．

インターネットが従来のメディアと根本的に異なるのは，サーチと双方向性である．従来のカタログ販売と最近のe-コマースを比較してみよう．カタログの場合，ページをめくりながら，購入したい品目があれば，そのページの端を折りこんでおく．Webサイトの場合は，サーチを利用する．どちらの方法も，運や偶然に左右されるものの，それまでは自分が欲しいとも気づかなかった品物を発見するのに役立つ．実際，サーチは，従来のメディアには太刀打ちできないほどの生産性をもたらす．広い意味でサーチといわれる機能を利用すると，数百万人分も生産性を向上させることができるし，最近のグローバル経済の拡大にも寄与するという意見があるが，これも納得できよう．そのような事例として，宅急便FedEx社の問合せ番号や調達システム，イントラネットがある．5年前なら試みようともしなかったのに，今では数秒で必要な情報を探し出して知ることができる．

サーチの将来を考えるとき，テキストサーチとデータベース技術の統合，分散リポジトリのサーチ，文脈情報の活用，物理世界との融合の4分野に注目しておきたい．これらの分野は，今後の数十年間，我々の日常生活にも影響を与えるだろう．

情報検索（IR）とデータベース　今日のサーチ技術は，情報検索と統計解析の研究から派生したものだが，主として関連性に基づくドキュメントのランク付けに力点を置いている．関連性を決定する要因はたくさんあるが，関連性とは元々が曖昧なものであり，基本的な決め方は似たりよったりである．つまり，情報の関連性を示すさまざまな要因を統計的に解析し，対象ドキュメントの関連性を効果的に予測する1つの数字を導き出すのである．

一方，データベースは，集合に対する操作として，結合とソートと絞り込みに重点を置いている．データベース技術を使ったサーチエンジンを作成することもできるが，これをうまくやるのは難しい．問題は，データベースにとって，構造化されていない曖昧なものはあまり得意でないことだ．データベースというのは，元々，銀行口座や従業員記録のように厳密に構造化されたデータを操作するように設計されている．しかしながら，実際に，世の中の興味深いデータはたいていデータベースの中にある．

原文："*When Everything Is Searchable*", *Communications of the ACM, Vol.44, No.3, pp.53-55* (*March 2001*)．著者より許可を得て翻訳

無統制とエントロピーを特徴とするインターネットにとってサーチが不可欠なのと同じ理由で，オフィス用のサーチエンジンが欲しいものだ．

したがって，このようなデータをサーチできるようにすること，もっと一般的な言い方にすると，データ構造に関する知識を取り込めるようにサーチ機能を進化させることが必要である．現状の確率論的手法と，データベースをフルに活用する構造化問合せ方式とを組み合わせる必要がある．このような複合手法の開発が難しいのは実にさまざまな理由があるが，最大の難題は，構造といっても非常にばらついており，時間の経過に伴って進化すること，さらに，「状態」や「給与」といった簡単な言葉の意味も統一されていないことである．これらの問題はXML（Extensible Markup Language）で解決できるという研究者もいるが，XMLは構造を記述する共通の方法を提案しているだけであって，構造そのものの意味を解釈するにはあまり役立たない．

分散レポジトリ　現世代のサーチ技術は分散情報源をほとんど扱えない．基本的には，Webを巡回して，遠隔地にあるすべての情報源からデータを収集し，中央にある1つのデータベースに持ち込んで，そこで問合せを処理する．これでうまくいく．だが，うまくいくのはWebページだけであって，興味深いデータの大半が格納されている大規模なリポジトリに対してはうまくいかない．情報アクセスの総合的な仕組みを考える上で，すべてのデータを1個所に集めるのではなく，問合せをデータのある場所にリアルタイムで転送するほうがよいケースもある．そのような例として，分散データベースの関数移送（function shipping），カリフォルニア大学バークレイ校のFFF（Federated Facts and Figures）サイト（http://fff.cs.berkeley.edu），Gnutellaのサーチ機能（註：有名なNapsterは，サーチ機能に関しては中央集中型）などがある．しかし，実際にこのような手法を使えるようにするには，非常に困った問題がいくつかある．

第1に，情報の可用性とアクセス時間の問題がある．現在でも，集中管理方式なら，すべての情報をいつでも利用できる状態にしておき，しかも，これまではほとんど無理な注文であった1秒の何分の1という高速で，必要な情報にアクセスできることを保証できる．しかし，分散資源の場合，すべての情報資源をいつでも利用でき，さらにアクセスもできる状態にしておくことは保証できない．仮にこれが可能になったとしても，高速性は期待できない．このような実態をエンドユーザに意識させないようにするということもできない．Napsterのユーザは，人気の高い音楽コンテンツのダウンロードの際にこの不便さを経験しているだろう．資源分散の究極の形態であるピアツーピア・システムにとって，これは根源的な問題である．

第2に，これよりもっと難しい信頼性の問題がある．分散サーチの場合，一般的にはそれほど必要とされない高レベルの信頼性が求められる．たとえば，スパムのようなケースがある．経済的な思惑（トラフィックを増やしたいなど）から，信頼性の低いコンテンツを掲載するサイトもある．また，ユーザが一連の分散データベースを信頼できたとしても，個々の問合せに対しては「どのデータベースを利用すればよいのか」，また「それぞれのデータベースからの回答の相対的な有意性をどう判断すればよいのか」といった基本的な問題が未解決のままである．

第3に，ブランドや評価など，社会的な関連性にかかわる問題がある．現在のサーチエンジンは，ブランドと評価の両方を少しずつ取り入れている．サーチエンジンでは，不審なサイトやページをブラックリストに載せ，「Wall Street Journal」や「Centers for Disease Control and Prevention」のように，よく知られている評価の高いサイト（ブランドサイト）を重視する設計になっている．また，対象となるページに張られているリンク数に注目して評価のランクを決める．これは，ほかの人が手間をかけてあるページにリンクを張ったということは，そのページはそれなりに良質だと考えられるからである．リンク数に基づいて評価レベルを測定することができる．将来，データ資源の評価は，もっと直接的で自動的な方式で行われるだろう．評価というものは，人によってさまざまなので，あくまで個人的なものであり，グローバルなものではない．したがって，短期的には，サイトの信頼性（認証）に重点を置き，評価は後回しになるだろう．

文脈の力　評価というのが個人的なものであること

から，関連性というものも個人的なものであり，誰にでも共通の基準はないことが明らかになった．「関連性」を一義的に定義できないということは，問題をさらに困難にする．問合せの意味は，誰がいつどこで問い合わせたのかという文脈に左右される．たとえば，問合せとして，たった2つの単語しか指定されていなかった場合，数十億ページの中から「ぴったり」のドキュメントを探さなければならないとしよう．これは実にすごいことだ．日常生活のどのようなやりとりでも，特定の文脈が暗黙に前提とされているのである．これを踏まえると，サーチにとっても，何らかの文脈情報が必要になるはずだ．サーチの文脈というのは，問合せをした人だけではなく，いつどこで問い合わせたのかという簡単なことについても知ることを意味する．たとえば，「よいレストランは？」という問合せは，誰がいつどこで問い合わせたのかという3つの文脈情報に依存する．同じ人物であっても，職場か家庭かで役割が異なるので，文脈が異なり，これに伴い，関連性の基準も異なる．

健康関連サイトなどの垂直ポータルは，文脈情報をある程度活用しているといえるだろう．また，さまざまな場所に移動中のユーザに対して，その現在位置を考慮して最寄りのレストランやお店を探し出すようなサービスも文脈情報の活用例になるだろう．しかし，一般には，サーチエンジンには，ユーザや問合せの目的などに関する役立つ情報がない．したがって，今まで，文脈情報に基づいて関連性を識別するというサーチエンジンの実践経験がほとんどない．今後は，CYCプロジェクト（http://www.cyc.com）のような役立つ形で暗黙の前提条件をコード化する必要がある．なお，CYCプロジェクトは，常識を構成する数千万の事実や発見的手法をコード化した先駆的な試みであった．サーチシステムに文脈を直接取り込むにはまだ数十年はかかるだろうが，もし実現すれば，すばらしい力を発揮するだろう．ユーザが異なれば，また同じユーザでも状況が異なれば，それぞれ違う結果が得られる．いずれにしても，現在のシステムが提供する「最適な」答えより，少しは妥当なものになるだろう．

物理世界との融合　長期的には，物理的な世界をもサーチの対象にする試みが行われるだろう．無統制とエントロピーを特徴とするインターネットにとってサーチが不可欠なのと同じ理由で，オフィス用のサーチエンジンが欲しいものだ．今日，我々は，紛失した荷物を探したり，どこかに置き忘れたコードレス電話を探したり，飛行機の場所を確認したりと，苦労している．しかし，仮想世界と物理世界の融合を推進すれば，サーチ力も強化・拡張できる．カリフォルニア大学バークレイ校とXerox PARCでの研究によると，元々回路をプリントするための高級なインクジェット・プリンタを使えば対話型のラベルを低コストで作成できるし，集積回路技術を使って小型のネットワーク型センサーとアクチュエータ（「スマート・ダスト」）を作成できるという．物理的な物の追跡用として，より詳細な説明のあるWebサイトへの案内ラベルとして，価格と機能の比較などの情報用として活用するのは，これらの極単純な用途の例である．さらに，本やファイルが本来の置き場所を自分自身で知ることができるようにもなる．一般化していえば，在庫管理がもっと簡単で強力になるということである．このようなサーチエンジンがオフィスでも使えるようになるだろう．

混沌に対する抜本的な解決策

無秩序で混沌とした情報に対する抜本的な解決策はサーチしかない．誰でもが情報の発信者になれ，情報と表現の自由を享受できるのも，サーチがあってこそ意義がある．我々は，このような無秩序な状態に直面し，信頼性と価値の相対化に関する根本的な問題を解決する第1歩を踏み出しただけにすぎない．同様に，従来のサーチ手法を，特許や国勢調査のデータなど，構造化された重要なデータが格納されている大規模リポジトリと組み合わせる方法についても，理解の端緒を開いたところである．長期的には，サーチ力を完全に発揮させられるかどうかは，文脈情報を活用し，インターネットと物理世界の統合を促進できるかどうかにかかっている．現在は，何世紀にもわたる情報共有の歩みのなかで最もおもしろく意味深い時期であり，将来もこの状態が引き続くことは間違いない．

謝辞　この翻訳では，南山大学の青山幹雄先生，日立教育部の田口昭仁氏から貴重なアドバイスをいただいた．最終的には訳者の判断で適宜採用させていただいた旨を明らかにし，各氏に感謝申し上げる．

（平成13年8月1日受付）