第14回

なぜ今、「ビッグデータ」か？

コンピュータの歴史はデータ容量増大の歴史

最近IT関係の記事で「ビッグデータ」という言葉をよく見かけます。実はこのビッグデータは、最近本欄で取り上げたIoT（モノのインターネット）やAI（人工知能）とも関係が深いのです。そこで本号ではビッグデータを取り上げることにします。

ビッグデータは、文字通り大量なデータを扱うものです。しかし、コンピュータの歴史はデータ量増大の歴史でもありました。

1960年代の初期の磁気ディスク装置の容量は数メガバイト（メガは100万）でしたが、現在家庭のパソコンで使われている外付けのハードディスクの容量は数テラバイト（テラは1兆）です。つまり、この50年間に容量が100万倍になったのです。そして、ディスクの進歩は現在も続いています。

ディスクの大容量化に伴って、その使われ方も大きく変わってきました。初期にはディスクが非常に高価だったため、格納されるデータは経営上の重要な数値や技術計算の結果などに限られていました。しかし現在は、パソコンのハードディスクに家族の写真やビデオ映像を大量に格納して楽しんでいるのです。

このように、コンピュータが取り扱うデータ量は長年月にわたって増大し続けてきたのです。ではなぜ最近、ことさらに「ビッグデータの時代来る！」と騒がれているのでしょうか？

ウェブの検索が道を切り開いた

1990年代にインターネットが急速に普及しました。そして、企業や個人が公開したい情報をウェブ記述言語で書いて全世界に流すようになりました。それを、ウェブ閲覧用のソフトであるブラウザで見ることができるのです。

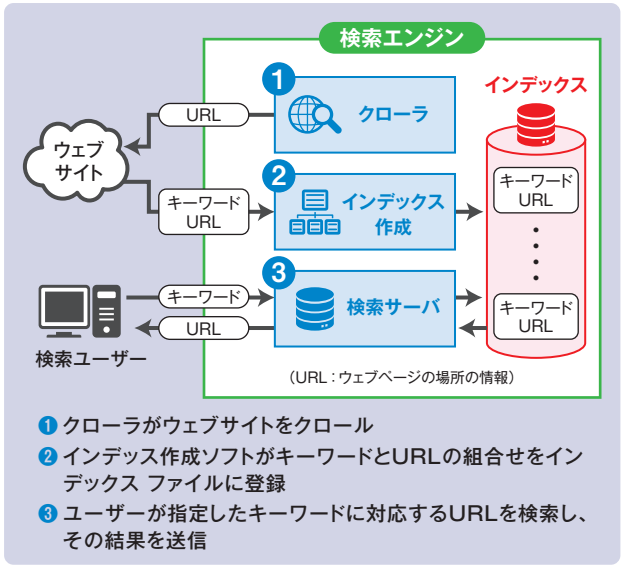
こうして、政府や企業の公開情報、個人の写真や絵などの作品、そして百科事典、辞書、地図などを、誰でもどこからでも閲覧できるようになりました。しかし、見たい情報がインターネット上のどこにあるかを捜すことが容易ではないことが問題でした。

そのため、初期には、ウェブサイトをカテゴリ別に分類して登録した「ディレクトリ」が主に使われていました。しかし、ウェブ情報が爆発的に増えると、こういう人手に頼った方法は通用しなくなりました。

そこで、「クローラ」という、全ウェブページを「クロール」（這い回ること。水泳のクロールと同じ）してキーワードを捜し、それをコンピュータが自動的に「インデックス」に登録する方法が考え出されました。書籍の索引（インデックス）と同じようなものを全ウェブページについて作るのです。この仕組みは「検索エンジン」と呼ばれ、現在全世界のウェブ検索の約3分の2にグーグルの検索エンジンが使われています。

こういう方法で、ユーザーが指定した複数のキーワードを含んだウェブページを短時間に探し出して表示するには、多数のサーバを同時に並列に使って、ユーザーの問合せを短時間に処理する必要があります。グーグルはそのために「グーグル・ファイル・システム」という特殊なファイル・システムを自社で開発し、数兆に及ぶウェブページから10京（京は兆の1万倍）バイトを超えるインデックスのファイルを構築しています。

グーグルはウェブサイトの検索に使っている機器やソフト、データセンタの場所などをほとんど公開していません。しかし、詳細は分からなくても、多数のサーバや記憶装置による並列処理で、従来考えられなかった規模のデータ処理がごく短時間でできることが示されました。



検索エンジンの仕組み

ビッグデータの道具が勢揃い

グーグルに代表されるウェブの検索によって、新しい情報処理の可能性が示されましたが、一般の企業にとって、グーグルのように自社でソフトを開発することは困難です。ところが、こういうソフトを無料で提供する団体が現れました。

たとえば、大規模なデータを高速に処理するには、多数のサーバによる並列処理が不可欠ですが、グーグルのそのソフトは非公開で他社は使えません。しかし、同様な処理を行う「Hadoop（ハドゥープ）」というソフトが無料で公開されました。「フェイスブック」や、前号で紹介した、クイズ番組で賞金王を破ったIBMの「ワトソン」はこのHadoopを使っています。

また、一般の企業がグーグルのように大規模なデータセンタを自前で持つことは困難です。しかし近年、大規模なセンタの設備を、使った分だけ費用を払うことによって利用できるクラウドというサービスが普及しました。そのため、必ずしも自社で設備を持つ必要はなくなりました。

こうして、中・小規模の企業でもビッグデータを扱うことができるようになったのです。

ビッグデータの技術を生かすためには・・・

こうして道具立てが揃ったので、いろいろな方面でビッグデータの技術が適用されるようになりました。IBMの「ワトソン」もその一例です。

しかし、Hadoopに代表されるビッグデータの技術を真に生かすことができる分野は限られています。

第1に、データ処理が多数のサーバで並列に行えることが必要です。ウェブの検索や「ワトソン」でのウィキペディアの検索がこれに当たります。いくらデータが大量でも、天気予報に使われる大気の状態の変化の数値計算などは、スーパーコンピュータの出番で、ビッグデータの技術は適しません。

第2に、処理結果を速く取得することに重要な価値があることです。「ワトソン」の早押し競争のクイズなどはまさにその例です。ほかにも、イベントの入場者の顔認識によるチェックなどがこれに当たります。これらの処理では、いかに正確な答えが得られても、長時間がかかったのでは意味がありません。

第3に、データ量が多いだけでなく、データ形式が統一されていることが重要です。

たとえば、医療情報で、患者の血圧、心拍数などの計測値、X線やMRIの画像、治療の経過などの情報を蓄積しておいて、類似症例を検索することによって診断や治療に役立てようとするシステムがあります。その際、データ形式が病院間でバラバラだと他の病院のデータは利用できません。データ形式が同じで、全病院のデータが活用できることが望まれます。

どの新技術についても言えることですが、ビッグデータも、その技術の特長が真に生きるところで使う必要があります。