10章 情報検索とデータベースの概要

データベースとは、「あるルールにしたがって蓄積されたデータや情報の集まり」をいう。ここでは、このデータベースの概念と役割、情報の集まりから主体的に目的にあった情報を収集することができる能力を育成する。

1. 目的とねらい

普通教科の「情報A」では「情報の収集・発信と情報機器の活用」、「情報B」では「モデル化とコンピュータを活用した解決」、「情報C」では「情報の収集・発信と個人の責任」において学習する。また、専門教科「情報」においては、ファイルとデータベース、データベースの仕組み、データベースの設計と操作などデータベースの概要についての内容を理解させる。
また、「データ通信の概要」のインターネットに関する内容、「アルゴリズムの基礎」のデータ構造などの知識とも関連する。

2. 情報とデータ

日常の生活において「この情報は確かか?」とか「あのデータが欲しい」などと、「データ」や「情報」という言葉をよく利用する。データ、情報、データベースは、それぞれ次のように分けられる。

(1)データとは

観測、測定、統計などから得られる客観的な事実で、文字、数字、画像、音声など人間が知覚できる形の表現方法で表わすことができるものである。さらに、人間または自動的な手段によって行われる通信、解釈、処理に適するように形成された事実、概念、指示などの表現である。
人間に関するデータというと、氏名・住所・学校名・病歴など文字データ、身長・体重・生年月日・成績などの数字データ、顔写真のような画像データとなる。

(2)情報とは

受け手が必要としている知識であり、受け手の価値観を変化させるデータで、ただ単に収集したデータと区別される。たとえば、ある状況で的確な判断を下したり、行動をとる意思決定において、データにある解釈を行なった結果、受け手の持つ知識を増加し、状況を変化させると、データは情報となる。
成績データを塾が入手すると、そこから弱点の教科を見つけ出し、特訓講座を受講するように勧める。そこで、成績データは、塾にとって情報となる。しかし、成績データを、レストランで使っても情報にはならない。

2.1 データベース

データベースは、ある一定のルールに従って蓄積されたデータや情報の集まりである。データベースは、次のような特徴を持つ。

(1)データベースを管理することができる

データベースは、蓄積されたデータを複数の人が参照したり、データの格納・更新・削除などの管理ができる。

(2)データベースをコンピュータで扱うことができる

データベースをコンピュータで管理すると、データの保管、検索・参照、格納・更新・削除ができる機能を持つ

(3)データベースはファイルの集まり

コンピュータで管理されるデータベースは、複数のファイルの集まりである。
企業では、製品情報、顧客情報、商品在庫情報、営業情報、人事情報、経理情報など多くのデータをファイルとして蓄積している。これらの情報は、企業活動にとって欠かせない重要な情報である。そこで、これらの情報を一括で管理し、必要な情報を引き出せるようデータベースを構築し、利用している。

このほか、パソコンで利用するアプリケーションソフトウェアには、「データベース機能」を備えていて、データベースの機能の一部を実現しているものもある。

2.2 データウェアハウス

データベースを構築して重要な情報を効果的に管理するとともに、これを企業全体で有効活用する中で考え出された技法である。企業内にあるデータベースのデータやインターネットなど外部から収集した情報を加えて、データを多次元的なデータベースに組み替える。これにより、企業の経営戦略の意思決定などに活用できる。さらに、データウェアハウスの高度な活用法として、大量のデータから法則、関連、知識などの隠れた情報を見つけ出す「データマイニング」という発見型の技法もある。
たとえば、商品を製造している企業には、商品の取引先と出荷量の変化しかわからないが、実際に商品を置いているお店で、別会社の商品とセットで買う人が多いという情報が入ると、そこで「別会社の商品」が足りないということがわかる。

2.3 データベースの活用方法

データベースを活用する1つの方法は、「情報検索」である。情報検索を行なうときは、情報が入っているデータベースを選択し、欲しい情報の条件を入力して探す。
最も身近な情報検索は、図書館での蔵書検索である。図書館で欲しい本を探すとき、いちばん簡単な方法は、本の棚を端から順に、背表紙を見て探す。しかし、この方法は時間がかかって効率的ではない。データベースを活用するときは、書名か著者名のどちらかの蔵書カードから検索する。蔵書カードは、五十音順などルールに従って並んでいる。目的の本の蔵書カードが見つかったら、蔵書カードに書いてある分類番号や棚番号を見て、その本棚から本を探すことができる。
しかし蔵書カードは、不便なことが多い。まずカードを引き出してしまうと、自動的に元の位置に戻らないし、戻すのが面倒である。また、本が貸し出されているかどうかは、貸し出しカードで確認しないとわからない。さらに、書名や著者名をはっきり覚えていないとか、本の内容や出版社しか覚えていない場合は、蔵書カードで検索することが困難になるか、時間がかかる。
そこで、これらをコンピュータのデータベースにしておくと、こういった不便なことが解消される。まずカードを見ても、カードが箱から消えることはない。コンピュータなら、出版社名、著者名や書名の一部だけ、本の内容といった項目でも検索できるように構築することができる。貸出管理システムと連携させれば、貸出中かどうかはすぐにわかる。

図10.1　図書館での本の検索風景

2.4 データベースでの検索

コンピュータで処理されるデータベースで情報を検索するときは、検索の手がかりとなる「キーワード」を使って、検索を行なう。
一般的に、「書名」や「著者名」などの項目をキーワードとしてそのものと一致するものだけを検索するものである。

(1) シソーラス

蔵書のようなデータベースでは、著者名を間違えると異なるものとなる場合はこれでよいが、外来語や概念などを伴う場合は、1つのものを同じ用語で読んでいるとは限らない。たとえば、コンピュータについて検索したいときは、「コンピュータ」「パソコン」「計算機」「情報端末」などもいっしょに検索できるとありがたいときがある。このように情報を提供する側は、同義語、上位語、下位語などキーワードとして関連するものを用意しておく必要がある。この関連する用語を整理して体系化し、情報検索をより効率的にするものを「シソーラス」という。

(2) 類似型文書抽出法

単語だけでなく、「××について○○が書いてある本」といった探し方をする方法がある。とくに専門書などでは、「栄養バランスを考えたお菓子の作り方の本」とか、「ペットとして飼える爬虫類の本」という風に探す。この場合は、コンピュータが入力された文章を解析し、名詞や動詞を探し出し、関連するキーワードとして検索する。これを「類似型文章抽出法」という。

2.5 インターネットでの検索

データが蓄積されているという意味では、インターネットは大きなデータベースの一つと考えられる。「検索エンジン」にキーワードを入力して、日本だけでなく、全世界のインターネットから欲しい情報を検索することができる。検索エンジンには、いくつかの種類があり、それにより結果が違ってくる。
たとえばYahoo!、Goo、Lycosなどが検索エンジンである。

(1)ロボット検索

インターネット上で見つけることができるWebサーバー上を定期的に回って情報を収集し、その情報の索引づけを自動で行なう。検索される件数は多いが、キーワードだけで検索するために、検索されたホームページにたまたま同じ文字の並びがあるだけで、内容が一致しない場合もある。つまり、無関係な情報が含まれる可能性が高い。

(2)ディレクトリ系検索エンジン

キーワードとの関連づけをするときに、その情報の要約に基づいて索引を付けるので、件数は少ないが、内容が一致する可能性が高くなる。また、生活、芸術、教育、趣味などの分野に分類されているので、これらの索引をたどって探すこともできる。

検索した結果が多い場合は、さらに「絞り込み検索」という方法を使う。検索された結果の中から、さらに関連するキーワードを入力して、情報を絞り込んでいく。

3. コンピュータとデータベース

蔵書検索でわかるように、データベースは、コンピュータで処理する情報としては最適なデータである。また、インターネットの普及でデータベースの利用が身近になった。そこで、コンピュータのデータベースに関してまとめる。

3.1 歴史

データベースがコンピュータで利用されるきっかけは、軍事関係であった。1950年代の米国国防省が、世界で展開している兵員や武器などの情報を一元管理するためにコンピュータを利用し、「情報(データ)の基地」という意味でデータベースと呼ばれるようになった。その後、この技術は民間企業でも利用されるようになり、データベースという言葉も一般化した。
日本においては、1950年代後半に、銀行業務にオフラインデータベースの会計処理システムが導入され、コンピュータの利用が本格的に始まった。1960年に入って当時の国鉄の「みどりの窓口」における座席予約システムが、オンラインデータベースとしての最初であった。
1970年代に入ると、情報通信ネットワークが発展し、コンピュータを相互に接続するコンピュータのダウンサイジングにより分散処理が発達し、データベースの需要も増加してきた。学校においても、成績、名簿、図書などのデータベース構築が盛んになった。

さらに1990年代に入って、パソコンの発達により、事務職や個人でも住所録などのデータベースを保有するようになった。さらに、インターネットの普及により、今では携帯端末などからでもデータベースを簡単に利用できるようになった。

3.2 データベースの分類

データベースを利用する立場からデータベースを分類すると、次のようになる。

(1) 情報の形態による分類

・ファクトデータベース：収集されたままの資料やデータを蓄積したもので、1次情報という。数値情報（統計データ、測定値など）、文字情報（文献、記事など）、画像情報（写真、地図など）、映像情報（自然現象、実験など）、音情報（音声、音楽など）などがある。
・リファレンスデータ：1次情報を加工したもので、2次情報という。1次情報を検索しやすくするために、分類したり、キーワードを付けたもの。文献情報をタイトル、発表者名、時期、ページ数などに分類したものなどがある。

(2)提供形態による分類

・オンライン：利用者の端末と情報を提供するデータベースが、通信回線で接続されているもので、リアルタイムに検索することができる。
・オフライン：情報を記憶したフロッピーディスクやCD-ROMを媒体として提供されるもの。文字のほかに、音声や画像も利用することができ、国語辞典や百科事典などがある。

(3)その他の分類

・データベースに蓄積されている情報の分野によって、自然科学技術、社会科学、医学、ビジネスなどに分類する方法。
・データベースの用途によって、一般に公開されて有料で利用できる商用データベースで、一般公開されていない。利用者を限定したデータベース。
商用データベースには、特許庁の日本特許・実用新案、新聞社が公開している新聞記事などがある。

3.3 データベースの仕組み

データベースは、情報を検索しやすように構築されていること、欲しい情報が適切なデータベースにあること、データの維持管理がしやすいことが重要である。そのために、データベース管理システム（DBMS：Data Base Management System）と呼ばれるシステムがある。これには次のような特徴がある。

(1) プログラムの独立性

データベースを実際に操作するのは、プログラムであり、このプログラムはデータと独立している。データベースに変更があっても、プログラムに影響がでないこと。

(2) データの重複はなし

データベースを一元管理するために、データに重複はないし、重複するデータを排除することができる。

(3) 同時処理の排除

データベースのアクセスは複数の人が同時に行なうが、書き込みや削除については、同時に行なえないように制御する。

(4) データの機密性

データベースのアクセスを制御することができ、アクセスできるユーザーを制限できる。

(5) データベースの障害回復

何らかの障害が発生したときに、これを回復する手段を準備していること。

3.4　データのモデル化とデータの構造

データベースを構築し、コンピュータに保管するときは、現実のデータをモデル化する。データベースを構築するときのもっとも基本的な項目として、現実のデータベースをどのように構造化し、コンピュータ内部で物理的にどのように表現するかである。次の3つの基本構造に分かれる。

(1) 概念モデル(概念スキーマ)

データを処理するうえで、コンピュータシステムに現実世界全体のデータを論理的に定義するもの。コンピュータやプログラムの特性を意識せずに、データそのものの立場で定義する。

(2) 外部モデル(外部スキーマ)

データベースを利用するプログラムの立場からデータベースを定義するもの。概念モデルで定義されたデータ構造の一部として考える。

(3) 内部モデル(内部スキーマ)

概念モデルで定義されたデータベースを、具体的にコンピュータ上にどのように表現するかを定義するもの。1つの概念モデルに1つの内部モデルが対応する。

また、データベースを構成する内部モデルでは、データ構造(各データがどのような関係で結びついているか)により、次のように分けることができる。

(1) ツリー(木)構造

データどうしの関係が親子関係のように、階層化されたデータベース。階層は、ツリー(木)構造で表現される。

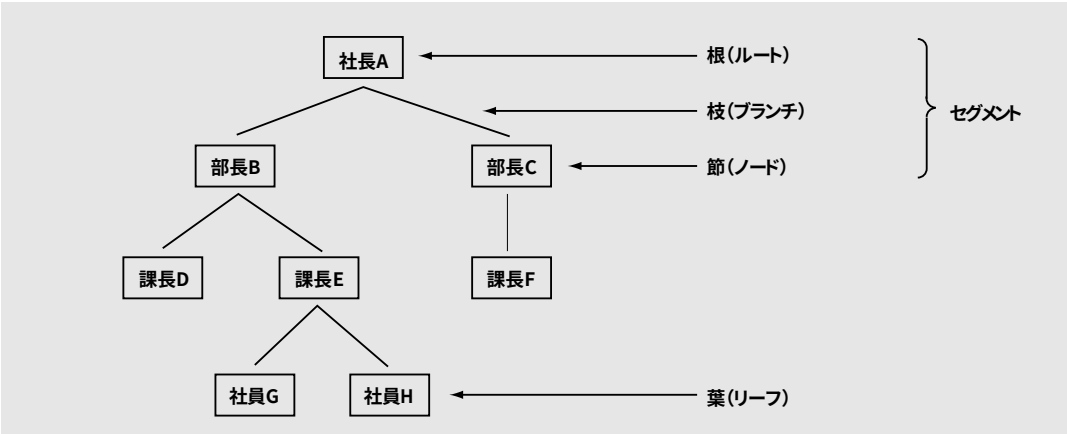

図10.2　ツリー構造

(2) ネットワーク(網)型

米国のデータシステム言語協議会(CODASYL)が提案した仕様に基づくデータベースで、データどうしの関係を網目状に設定したもの。1970年代から導入されたホストコンピュータで管理される企業の大規模データベースは、この型が多い。

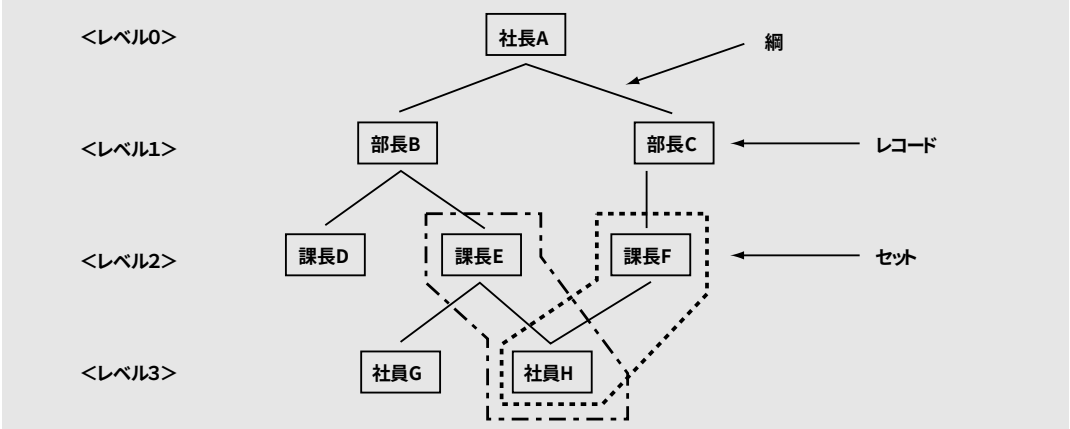

図10.3　ネットワーク型

(3)リレーショナル型

1980年代に米国IBMの研究者(E.F.Codd)によって提案されたデータベースで、データベース内のデータをいくつかの2次元の表によって表現するデータベース。構造がシンプルで、データの組み合わせが自由であり、現在のパソコンやサーバー・クライアント型システムなどでのデータベースの主流となっている。

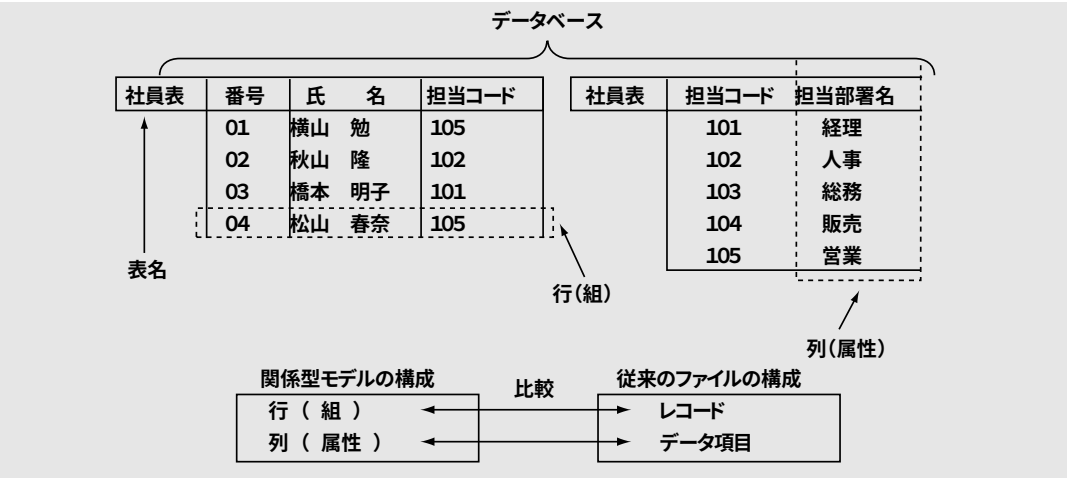

図10.4　リレーショナル型

3.5　データベースの設計と構築

データベースを導入するには、情報をいかにデータベース上へ反映させるかが重要である。そこで、データベースの設計がポイントとなる
データベースの設計は、概念スキーマを設計する「概念設計」、概念スキーマをもとに特定のDBMSに対応した外部スキーマと内部スキーマを作成する「論理スキーマ」、特定のDBMSが提供する物理的な実装レベルの構成を記述した内部スキーマを設計する「物理設計」の段階がある。
この内部スキーマを記述するのが、データベース言語である。リレーショナル型では、SQL(Structured Query Language)と呼ばれるデータベースがある。

4.　指導のポイント

授業においては、講義だけにならないよう、インターネットを利用して情報を収集したり、収集したデータを読み取るなど、コンピュータとネットワークをできるだけ利用する。ただし、深入りすることなく、概要説明にとどめる。また、各データベースについて、理解を深めるため、活用方法やメリットやデメリットなどを考えさせることが大切である。

【演習例】
(1)学内で収集できるデータで、データベースとすると最適なものを考える
(2)インターネットを活用して、データベースサービスを利用する
(3)インターネットで情報検索を行ない、検索の実際を体験する