第 1 章

基 本 概 念

本章首先概述数据管理的发展历史，其次介绍度量空间的基本概念，最后给出目前度量空间数据管理面临的挑战。

1.1 数据管理

自 20 世纪 60 年代以来，数据从单一型不断演变成复杂多样型。相应地，数据管理系统也逐渐从定制型系统逐渐演变成通用型系统。图 1-1 描述了数据管理系统的发展历史。

1.1.1 关系型数据管理系统

在 20 世纪 60 年代初，随着计算机在企业管理中的普及，部分大公司开始建立自己的信息管理系统，用于记录员工 ID、产品价格等数值信息。因而产生了关系型数据库系统，用于管理一维数据，并支持数据选择和连接等基本操作。20 世纪 70 年代，研究人员提出了 B^+ 树索引以提高关系型数据管理系统的查询效率。

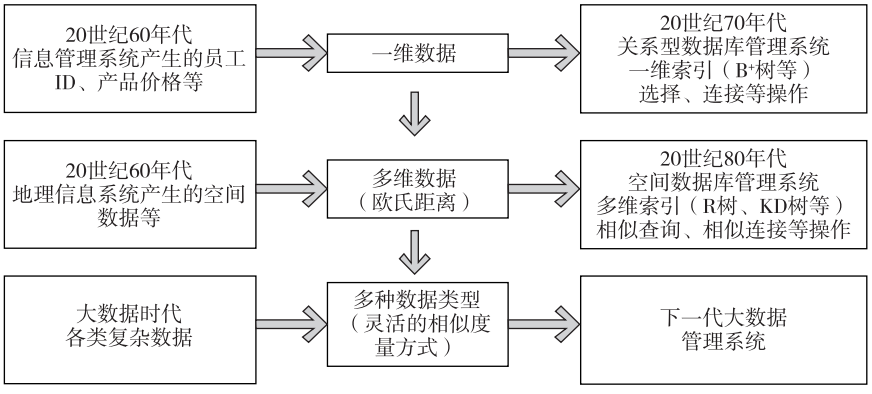

图 1-1　数据管理系统发展历史

1.1.2　空间数据管理系统

随着人造卫星的发明，地理空间信息可以通过卫星获取，地理信息系统应运而生。因此，研究人员开发了空间数据库管理系统以有效管理空间数据，其中，空间数据可以通过多维特征向量来表示，且空间数据间的相似性可以使用欧氏距离来度量。自 20 世纪 80 年代以来，研究人员提出了大量空间索引(如 R 树、KD 树等)，用于提升空间数据库管理系统的查询效率，并探讨了多种空间查询(如相似查询等)以支持多样的地理信息系统应用。

1.1.3　下一代数据管理系统

随着计算机、互联网、通信以及定位技术的快速发展，数据量呈爆炸式增长，导致大数据时代的降临。为此，需要开发通用的下一代大数据管理系统以有效地管理大数据。

大数据通常具有以下四大特性(即 4 个 V)：数据量大(Volume)、数据类型繁多(Variety)、价值密度低(Value)以及更新速度快(Velocity)。针对数据类型繁多的特性，本书采用了一个通用的数据表达模型(即度量空间)，以便对多源异构大数据进行有效建模；针对数据量大、价值密度低和更新速度快的特性，本书介绍了度量空间数据管理技术，以支持高效的多源异构大数据分析。

1.2 度量空间

从传感网和物联网中各种遥感、测绘以及监控设备采集的巨量实时流媒体数据；从互联网公司的各类服务中获得的海量文档、通信、新闻、网页、交易等数据；从基因组学、蛋白组学、脑科学和天体物理学产生的数以亿计的医学、生物、气象、地质数据。这些数据类型复杂多样(如图 1-2 所示)，且数据之间的相似性并不能用欧氏距离来简单度量，而需要用其他的距离度量方式，如网络距离需要借助于最短路径，文本间的距离需要借助于编辑距离等。因此，针对大数据的数据类型繁多的特性，我们需要一种通用的数据表达模型，以支持各种数据类型和多样的相似度量方式。

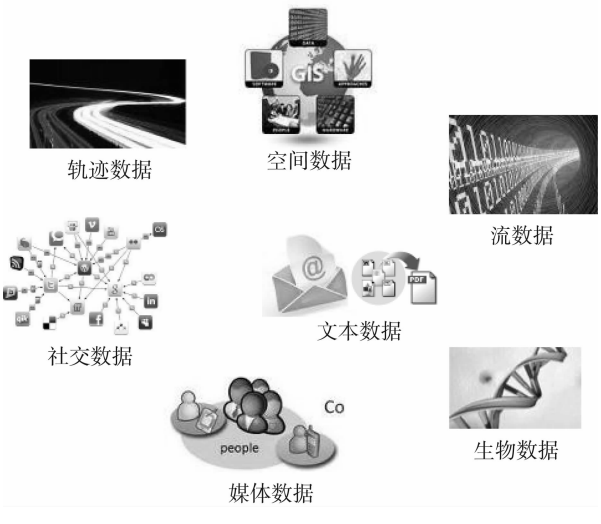

图 1-2 多类型数据

度量空间是一个(M, d)二元组，其中 M 是一个数据对象集，d:$M\times M\rightarrow \boldsymbol{R}$ 是一个距离函数。M 中的数据对象不一定包含坐标信息，它可以表示任意的数据对象(如位置信息、DNA 序列、时间序列、颜色直方图等)；距离函数 d 给出了数据对象之间的相似度量方式，并具有 5 个特性：

1)对称性(symmetry)：$\forall x, y \in M, d(x, y) = d(y, x)$。

2)自反性(reflexivity)：$\forall x \in D, d(x, x) = 0$。

3)非负性(non negativity)：$\forall x, y \in M, d(x, y) \geqslant 0$。

4)严格非负性(positiveness)：$\forall x, y \in M, x \neq y \Rightarrow d(x, y) > 0$。

5)三角不等性(triangle inequality)：$\forall x, y, z \in M, d(x, z) \leqslant d(x, y) + d(y, z)$。

度量空间支持任意满足上述 5 个特性的距离度量方式，如 L_p 范数(L_p-norm)、编辑距离和最短路径等。根据度量空间的定义，度量空间下的数据不受几何特性限制，且数据对象之间的距离度量方式只需满足上述 5 种特性即可。因而，度量空间可以作为大数据的一种通用表达模型，可以在一定程度上解决高维数据和多类型数据问题。因此，度量空间下的索引与查询技术研究具有重要的理论意义和广阔的应用前景，其不仅可以丰富度量空间领域的研究成果，而且还可以促进数据挖掘等相关领域的进一步发展。

1.3 存在的问题

尽管目前已开展一定的度量空间数据管理技术研究，但仍面临以下 3 大挑战。

(1)缺乏高效的索引结构

索引是加速查询的一种有效方式。现有的度量空间索引可以分为基于支枢点的索引方法、基于划分的索引方法和混合索引方法。基于支枢点的索引方法在距离计算次数(即 CPU 代价)方面优于基于划分的索引方法，但其存储空间消耗过大且 I/O 代价较大。此外，已有的混合索引方法研究尚少，且该类方法的存储空间消耗依然很大。另外，在现实生活中，设备的局限性、持续的数据更新、隐私保护和高通量测序技术等可能导致数据的不确定性。因此，如何设计高效的索引结构是度量空间数据管理研究的一大关键问题。

(2)查询复杂度高

度量全 k 最近邻查询和度量 k 最近对查询涉及两个数据集的连接操作，因而查询复杂度高。现有的度量全 k 最近邻查询算法针对每个查询

点进行一次度量 k 最近邻查询，所以，对于查询集较大的情况，现有的算法存在大量冗余的 I/O 和 CPU 开销。为此，如何降低度量空间查询复杂度是度量空间数据管理研究的又一关键问题。

(3)缺乏查询可用性分析

现有的度量空间查询(如度量概率区域查询)研究旨在提升查询性能和减少存储空间，而缺少对查询结果可用性的关注。例如，一个用户进行了一次度量概率区域查询，当其得到查询结果时，他/她发现其真正需要的对象却没有出现在查询结果中。此时，该用户可能考虑为什么他/她期望的查询结果没有出现，如何调整才能得到期望的查询结果。因此如何有效地提高用户对查询结果的满意度(即提高查询可用性)也是度量空间数据管理研究的一大关键问题。

针对上述度量空间数据管理研究存在的挑战，本书主要介绍了度量空间索引、集中式度量空间查询、分布式度量空间查询和度量空间查询可用性分析。