

**陈** **明**◎编著

大数据

基础与应用

◆大数据技术与教育**扛鼎之作**

◆全面系统深入讲述**大数据**

北京师范大学出版集团

BEIJING NORMAL UNIVERSITY PUBLISHING GROUP

址 京 师 范 大 学 出 版 社



**陈明◎编著**





 ]

**基础与应用**

北京师范大学出版集团

BEUING NORAL UNIVERSITY PUBLISHING GROUP

北京师范大学出版社



**图书在版编目** **(CIP) 数据**

大数据基础与应用/陈明编著.—北京：北京师范大学出版社，2016 ISBN 978-7-303-20018-4

I.① 大… Ⅱ.①陈… Ⅲ.①数据处理 IV.①TP274

中国版本图书馆 CIP数据核字(2016)第005840号

营销中心电话010 - 62978190 62979006

北师大出版社科技与经管分社网 hitp:/jsws.bnupg.com

电 子 信 箱k j i g @ b n u p g . c o m

出版发行：北京师范大学出版社 www.bnupg.com

北京新街口外大街19号

邮政编码：100875

印 刷：北京京师印务有限公司

经 销：全国新华书店

开 本：184mm×260mm 1/16

印 张：21.25

字 数： 504千字

版 次： 2016年3月第1版

印 次： 2016年3月第1次印刷

定 价：4 2 . 0 0 元

|  |  |
| --- | --- |
| 策划编辑：李 丹 | 责任编辑：李 丹 |
| 美术编辑：刘 超 | 装帧设计：刘 超 |
| 责任校对：马军令 | 责任印制：曲利华 |

**版** **权** **所** **有** **侵** **权** **必** **究**

反盗版、反侵权举报电话：010-62978190

北京读者服务部电话：010-62979006-8021

外埠邮购电话： 010-62978190

本书如有印装质量问题，请与印制管理部联系调换。

印制管理部电话：010-62979006-8006

**前** **言**

需求是科学技术发展的原动力。大数据问题的出现与研究已经成为计算机科学与技术 研究的新热点，并显示出日益强大的吸引力，科学大数据的出现催生了数据密集型知识发 现的第四科学研究范式的出现。对于信息领域，大数据带来的不仅是机遇，还有一系列的 困难和挑战。大数据技术与应用展现出锐不可当的强大生命力，科学界与企业界对其寄予 无比的厚望。大数据成为继20世纪末、21 世纪初互联网蓬勃发展以来的又一轮IT 工业革 命。数据本身是无意义的，而通过统计、分类、萃取、特征抽取等一系列技术手段，可以 从数据中提取信息与知识。所以说，数据是重要的战略资源，隐含巨大的经济价值，已经 引起科技界和企业界的高度重视。有效地组织和使用数据，将对经济发展产生巨大的推动 作用。大数据是以大样本或全样本代替抽样，以近似代替准确，以联系代替因果，因此， 大数据是对传统的IT 各领域的挑战，研究大数据意义深远。大数据孕育着前所未有的机遇。 对大数据的交换、整合和分析，可以发现新的知识，创造新的价值，带来大知识、大科技、 大利润和大发展。

本书概括性地介绍了大数据的主要内容，是大数据技术入门的参考书。全书分为17章， 其中第1章是对大数据的简单概述，第2章介绍大数据研究的方法论，第3章、第8章、 第9章和第14章介绍大数据的生态环境，第4章、第5章、第6章、第7章、第10章、 第11章、第12章、第13章、第15章和第16章介绍大数据技术及应用方法，第17章简 单介绍数据科学的内容。

大数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，对数据的占 有和控制将成为陆权、海权、空权之外的另一种国家核心资产。联合国在2012年发布了《大 数据政务白皮书》,指出大数据对于联合国和各国政府是一个历史性的机遇，通过使用极 为丰富的数据资源，对社会经济进行前所未有的实时分析，帮助政府更好地调整社会和经 济运行。数据为王的大数据时代已经到来，对数据的占有和控制也将成为新的争夺点。大 数据技术的专业人才，特别是数据分析复合型人才的稀缺将会影响市场的发展。

在技术上，大数据、海量数据与超大规模数据都是指用传统处理方法无法处理的大量 数据。通过对大数据进行高速有效的处理，可以发现数据中蕴藏的规律，进而为各种关键 决策提供依据与指导。正确的预测与决策将导致巨大财富的产生。技术与工具密不可分， 目前常用的数据处理技术与工具适用于小数据处理， 一些海量数据处理方法与工具是过渡 性的，大数据处理技术与工具的研究是一项有理论意义和实际价值的工作。大数据的出现， 对 IT 各领域的传统处理方法提出了新的冲击与挑战。大数据技术是一门实践性较强的技 术，需要重视工具与应用方法的选择与研究。

**大数据基础与应用**

高校是培养人才的重要基地，需要分析和定位大数据带来的影响。大数据推动学科发 展和学科建设、利用大数据技术整合现有的教学资源为已经到来的“数据革命”培养行业 紧缺人才都是教育工作面临和需要解决的问题。

本书在结构上呈积木状，对各章内容进行独立的概念性论述。由于作者水平有限，书 中不足之处在所难免，敬请读者批评指正。



2015年10月



**目** **录**

**第** **1** **章** **走进大数据时代** [1](#bookmark2)

1.1 应对大数据 [2](#bookmark3)

1.1.1 电子数据迅速增加 [2](#bookmark4)

1.1.2 数据中蕴含的价值 [2](#bookmark5)

1.1.3 数据是国家的核心资产 [3](#bookmark6)

1.2 大数据的生态环境 [3](#bookmark7)

1.2.1 互联网世界 [4](#bookmark8)

1.2.2 物理世界 [5](#bookmark9)

1.3 大数据的概念 [6](#bookmark1)

1.3.1 数据容量 [6](#bookmark1)

1.3.2 数据类型 [8](#bookmark10)

1.3.3 价值密度 [8](#bookmark11)

1.3.4 速 度 [8](#bookmark12)

1.3.5 真实性 [8](#bookmark13)

1.4 大数据的性质 [8](#bookmark14)

1.4.1 非结构性 [8](#bookmark15)

1.4.2 不完备性 [9](#bookmark16)

1.4.3 时效性 [9](#bookmark17)

1.4.4 安全性 [10](#bookmark18)

1.4.5 可靠性 [10](#bookmark19)

1.5 大数据技术概述 [10](#bookmark20)

1.5.1 大数据处理的全过程 [11](#bookmark21)

1.5.2 大数据技术的特征 [13](#bookmark22)

1.5.3 大数据的关键问题与

关键技术 [14](#bookmark23)

1.6 大数据应用 [16](#bookmark24)

1.6.1 大数据应用趋势 [16](#bookmark25)

1.6.2 大数据应用评价与应用实例 17

小结 [21](#bookmark26)

**第** **2** **章** **科学研究范式** [22](#bookmark27)

2.1 科学研究第一范式 [23](#bookmark28)

2.1.1 科学实验特点与步骤 [23](#bookmark29)

2.1.2 科学实验构成与分类 [24](#bookmark30)

2.1.3 科学实验程序 [25](#bookmark31)

2.1.4 科学实验使用原则 [25](#bookmark32)

2.2 科学研究第二范式 [26](#bookmark33)

2.2.1 科学理论的特征与价值 [27](#bookmark34)

2.2.2 科学理论的结构与体系建立

方法 [27](#bookmark35)

2.3 科学研究第三范式 [28](#bookmark36)

2.3.1 系统模拟发展过程 [29](#bookmark37)

2.3.2 系统模拟基本方法 [29](#bookmark38)

2.3.3 系统模拟语言 [30](#bookmark39)

2.4 科学研究第四范式 [30](#bookmark40)

2.4.1 数据密集型计算 [31](#bookmark41)

2.4.2 格雷法则 [32](#bookmark42)

2.4.3 核心内容 [35](#bookmark43)

小结 [36](#bookmark44)

**第** **3** **章** **分布系统设计的CAP 理论** **37**

3.1 分布式系统的伸缩性 [38](#bookmark45)

3.1.1 可伸缩性的概念 [38](#bookmark46)

3.1.2 影响横向扩展的主要因素 39

3.2 横向扩展方案 [42](#bookmark47)

3.2.1 可伸缩共享数据库 [42](#bookmark48)

3.2.2 对等复制的横向扩展方案 43

3.2.3 链接服务器和分布式查询 44

3.2.4 分布式分区视图 [45](#bookmark49)

3.2.5 数据依赖型路由的横向扩展 46

3.3 CAP 理论 [47](#bookmark50)

3.3.1 分布系统设计的核心

系统需求 [47](#bookmark51)

3.3.2 CAP 定理 [49](#bookmark52)



大数据基础与应用

3.4 BASE 模 型 [53](#bookmark53)

3.4.1 三个核心需求分析 [53](#bookmark54)

3.4.2 ACID 、BASE 与 C A P 的

关系 [54](#bookmark55)

3.4.3 CAP 与延迟 [55](#bookmark56)

3.4.4 CAP 理论的进一步研究 [55](#bookmark57)

3.5 Web 分布式系统设计 [57](#bookmark58)

3.5.1 系统核心需求 [57](#bookmark59)

3.5.2 系统服务 [58](#bookmark60)

3.5.3 冗 余 [59](#bookmark61)

3.5.4 分区 [60](#bookmark62)

小结 [61](#bookmark63)

**第** **4** **章** **大数据网络空间** [62](#bookmark64)

4.1 复杂网络空间概述 [63](#bookmark65)

4.1.1 复杂网络概念与特征 [63](#bookmark66)

4.1.2 复杂网络的特性 [64](#bookmark67)

4.2 社会网络 [65](#bookmark68)

4.2.1 社会网络结构 [65](#bookmark69)

4.2.2 社会网络理论 [66](#bookmark70)

4.2.3 社会计算 [67](#bookmark71)

4.2.4 社会网络应用 [68](#bookmark72)

4.3 社会网络分析 [69](#bookmark73)

4.3.1 社会网络分析概述 [70](#bookmark74)

4.3.2 社会网络分析的原理 [70](#bookmark75)

4.3.3 社会网络分析的特征 [71](#bookmark76)

4.3.4 社会网络分析的方法 [71](#bookmark77)

4.4 社会网络中的隐私保护 [72](#bookmark78)

4.4.1 用户隐私类型 [72](#bookmark79)

4.4.2 身份隐私攻击与保护 [73](#bookmark80)

4.4.3 用户关系的攻击及保护 [73](#bookmark81)

4.4.4 万维网用户隐私保护 [74](#bookmark82)

4.5 社会感知计算 [74](#bookmark83)

4.5.1 社会感知计算概念 [74](#bookmark84)

4.5.2 社会感知计算的内容 [75](#bookmark85)

4.6 人类通信方式 [76](#bookmark86)

4.6.1 通信方式的演化 [76](#bookmark87)

4.6.2 六度分隔理论 [77](#bookmark88)



4.6.3 150法 则 [78](#bookmark89)

4.6.4 唯象理论与唯象方法 [79](#bookmark90)

4.7 社 交 网 站 [80](#bookmark91)

4.7.1 社交网站作用 [80](#bookmark92)

4.7.2 Web 网 站 [80](#bookmark93)

小结 [81](#bookmark94)

**第** **5** **章** **MapReduce 分布编程模型** **82**

5.1 函数式编程范式 [82](#bookmark95)

5.1.1 函数型语言与函数式编程 83

5.1.2 函数式编程优点 [83](#bookmark96)

5.1.3 函数式编程的特征 [84](#bookmark97)

5.2 映射函数与化简函数 [84](#bookmark98)

5.2.1 映射与映射函数 [84](#bookmark99)

5.2.2 化简与化简函数 [85](#bookmark100)

5.3 MapReduce 的体系结构 [86](#bookmark101)

5.3.1 MapReduce 计算描述 [86](#bookmark102)

5.3.2 MapReduce 适用情况 [88](#bookmark103)

5.4 基于Hadoop 平台的分布式计算 [88](#bookmark104)

5.4.1 Hadoop 发展历程 [88](#bookmark105)

5.4.2 分布式系统与 Hadoop [90](#bookmark106)

5.4.3 SQL 数据库和 Hadoop [90](#bookmark107)

5.4.4 基于Hadoop 的分布式计算 [92](#bookmark108)

小结 [98](#bookmark109)

**第** **6** **章** **大数据流式计算** [99](#bookmark110)

6.1 流式数据的概念与特征 [99](#bookmark111)

6.1.1 流式数据的概念 [99](#bookmark112)

6.1.2 流式数据的特征 [100](#bookmark113)

6.2 大数据的计算模式 [101](#bookmark114)

6.2.1 大数据批量计算模型 [101](#bookmark115)

6.2.2 大数据流式计算模型 [101](#bookmark116)

6.2.3 大数据流式计算与批量

计算的比较 [103](#bookmark117)

6.3 流式大数据处理工具 [104](#bookmark118)

6.3.1 Storm 系统 [104](#bookmark119)

6.3.2 S4系 统 [107](#bookmark120)

6.3.3 Data Freeway and Puma

系统 [110](#bookmark121)

6.4 大数据流式计算的应用 [111](#bookmark122)

6.4.1 金融银行业的应用 [112](#bookmark123)

6.4.2 互联网领域的应用 [112](#bookmark124)

6.4.3 物联网领域的应用 [113](#bookmark125)

6.4.4 三种典型应用场景的对比 113

小结 [114](#bookmark126)

**第** **7** **章** **大数据搜索技术** [115](#bookmark127)

7.1 搜索引擎概述 [116](#bookmark128)

7.1.1 搜索引擎的发展过程 [116](#bookmark129)

7.1.2 搜索引擎的定义 [117](#bookmark130)

7.1.3 搜索引擎的组成 [117](#bookmark131)

7.1.4 搜索引擎的分类 [117](#bookmark132)

7.1.5 搜索引擎的工作过程 [120](#bookmark133)

7.1.6 搜索引擎的评价指标 [121](#bookmark134)

7.2 语义搜索引擎 [121](#bookmark135)

7.2.1 语义与语义搜索引擎的

概念 [121](#bookmark136)

7.2.2 图谱 [122](#bookmark137)

7.2.3 搜索就是回答 [123](#bookmark138)

7.2.4 语义搜索引擎的组成 [123](#bookmark139)

7.2.5 基于本体的语义搜索引擎 123

7.3 网站数据对搜索引擎的影响 126

7.3.1 垂直网站与垂直搜索 [126](#bookmark140)

7.3.2 私有化的Web 化数据 [127](#bookmark141)

7.3.3 没有 Web 化的数据 [127](#bookmark142)

7.3.4 大数据流动 [128](#bookmark143)

7.4 搜索引擎优化 [128](#bookmark144)

7.4.1 搜索引擎优化的产生 [129](#bookmark145)

7.4.2 网页级别 [129](#bookmark146)

7.4.3 搜索引擎优化的方法 [129](#bookmark147)

小结 [131](#bookmark148)

**第** **8** **章** **大数据存储** [132](#bookmark149)

8.1 大数据存储概述 [132](#bookmark150)

8.1.1 大数据存储模型 [133](#bookmark151)

8.1.2 大数据存储问题 [133](#bookmark152)

8.1.3 存储方式 [135](#bookmark153)

8.2 大数据的存储技术 [139](#bookmark154)



目 录

8.2.1 数据容量问题 [139](#bookmark155)

8.2.2 大图数据 [140](#bookmark156)

8.2.3 分布式存储的架构 [142](#bookmark157)

8.2.4 数据存储管理 [143](#bookmark158)

8.3 数据云存储 [145](#bookmark159)

8.3.1 云存储的意义与问题 [145](#bookmark160)

8.3.2 技术措施 [146](#bookmark161)

8.4 数据存储的可靠性 [148](#bookmark162)

8.4.1 磁盘与磁盘阵列的可靠性 148

8.4.2 文件系统的可靠性 [151](#bookmark163)

小结 [151](#bookmark164)

**第** **9** **章** **NoSQL数据库** [152](#bookmark165)

9.1 NoSQL 概述 [153](#bookmark166)

9.1.1 非结构化问题 [153](#bookmark167)

9.1.2 NoSQL 的产生 [153](#bookmark168)

9.2 NoSQL 的特点与问题 [155](#bookmark169)

9.2.1 NoSQL 的特点 [155](#bookmark170)

9.2.2 NoSQL 面对的问题 [156](#bookmark171)

9.3 NoSQL 的主要存储方式 [157](#bookmark172)

9.3.1 键值存储方式 [157](#bookmark173)

9.3.2 文档存储方式 [158](#bookmark174)

9.3.3 列存储方式 [160](#bookmark175)

9.3.4 图形存储方式 [166](#bookmark176)

9.3.5 存储类型对应的 NoSQL

数据库 [167](#bookmark177)

9.4 常用的NoSQL 数据库 [168](#bookmark178)

9.4.1 Cassandra [168](#bookmark179)

9.4.2 Lucene [168](#bookmark180)

9.4.3 Riak [169](#bookmark181)

9.4.4 CouchDB [169](#bookmark182)

9.4.5 Neo4j [169](#bookmark183)

9.4.6 Oracle 的 NoSQL [169](#bookmark184)

9.4.7 Hadoop 的 HBase [170](#bookmark185)

9.4.8 Bigtable [170](#bookmark186)

9.4.9 DynamoDB [170](#bookmark187)

9.4.10 MongoDB [170](#bookmark188)

小结 [173](#bookmark189)





**大数据基础与应用**

**第10章** **大数据预处理技术** [**174**](#bookmark190)

10.1 数据抽取概述 [175](#bookmark191)

10.1.1 数据抽取的概念与包装器 175

10.1.2 抽取数据的方法 [176](#bookmark192)

10.2 Web 数据抽取 [177](#bookmark193)

10.2.1 Web 数据抽取问题的

提出 [177](#bookmark194)

10.2.2 Web 数据抽取的目的与

方法 [177](#bookmark195)

10.2.3 Web 数据抽取过程 [178](#bookmark196)

10.3 数据质量与数据清洗 [179](#bookmark197)

10.3.1 数据质量 [179](#bookmark198)

10.3.2 数据清洗的目的 [180](#bookmark199)

10.3.3 数据清洗算法衡量标准 180

10.3.4 数据清洗的主要研究成果 [181](#bookmark200)

10.3.5 数据清洗技术面临的

问 题 [181](#bookmark201)

10.4 不符合要求的数据 [182](#bookmark202)

10.4.1 不完整的数据 [182](#bookmark203)

10.4.2 异常的数据 [182](#bookmark204)

10.4.3 重复的数据 [183](#bookmark205)

10.5 数据清洗技术的实现 [183](#bookmark206)

10.5.1 数据清洗的方法与技巧 184

10.5.2 数据清洗的实现方式 [186](#bookmark207)

10.5.3 数据清洗的步骤 [187](#bookmark208)

10.5.4 数据清洗的评价标准 [188](#bookmark209)

10.5.5 常用的数据清洗算法 [188](#bookmark210)

10.5.6 大数据清洗工具 [189](#bookmark211)

10.6 数据集成 [189](#bookmark212)

10.6.1 数据集成技术概述 [190](#bookmark213)

10.6.2 数据集成系统的构建 [192](#bookmark214)

10.7 数据转换与约简 [193](#bookmark215)

10.7.1 数据转换 [193](#bookmark216)

10.7.2 数据约简 [194](#bookmark217)

小结 [195](#bookmark218)

**第** **1** **1** **章** **大数据分析** [196](#bookmark219)

11.1 数据分析概述 [197](#bookmark220)



11.1.1 数据分析的概念 [197](#bookmark221)

11.1.2 数据分析的目的与意义 197

11.1.3 数据分析方法的分类 [198](#bookmark222)

11.1.4 数据分析的类型 [198](#bookmark223)

11.1.5 数据分析步骤 [199](#bookmark224)

11.2 基本数据分析方法 [199](#bookmark225)

11.2.1 统计方法 [199](#bookmark226)

11.2.2 指标对比分析法 [200](#bookmark227)

11.2.3 分组分析法 [201](#bookmark228)

11.2.4 综合评价分析法 [201](#bookmark229)

11.2.5 指数分析法 [201](#bookmark230)

11.2.6 平衡分析法 [201](#bookmark231)

11.2.7 平滑和滤波 [202](#bookmark232)

11.2.8 基线与峰值 [202](#bookmark233)

11.3 高级数据分析方法 [202](#bookmark234)

11.3.1 时间数列及动态分析法 202

11.3.2 相关分析 [203](#bookmark235)

11.3.3 回归分析 [203](#bookmark236)

11.3.4 判别分析 [204](#bookmark237)

11.3.5 对应分析 [204](#bookmark238)

11.3.6 预测分析 [204](#bookmark239)

11.3.7 主成分分析 [204](#bookmark240)

11.3.8 多维尺度分析 [205](#bookmark241)

11.3.9 因子分析 [205](#bookmark242)

11.3.10 方差分析 [205](#bookmark243)

11.4 复合技术分析 [205](#bookmark244)

11.4.1 快速傅里叶变换 [206](#bookmark245)

11.4.2 分类 [206](#bookmark246)

11.4.3 聚类分析 [206](#bookmark247)

11.5 大数据分析基础 [207](#bookmark248)

11.5.1 可视化分析 [207](#bookmark249)

11.5.2 数据挖掘 [207](#bookmark250)

11.5.3 预测性分析 [208](#bookmark251)

11.5.4 语义引擎 [208](#bookmark252)

11.5.5 数据质量和数据管理 [208](#bookmark253)

11.5.6 大数据的离线与在线分析 208

11.6 大数据预测分析 [209](#bookmark254)

11.6.1 预测学简介 [209](#bookmark255)

11.6.2 预测原理 [211](#bookmark256)

11.6.3 预测的步骤 [212](#bookmark257)

11.6.4 预测技术分类 [213](#bookmark258)

11.6.5 预测模型及分类 [215](#bookmark259)

11.6.6 大数据预测分析要素 [215](#bookmark260)

11.6.7 大数据预测分析的演化 216

11.6.8 大数据预测分析相关问题 216

11.6.9 舆情监测与分析 [217](#bookmark261)

11.7 大数据分析应用 [220](#bookmark262)

11.7.1 为客户提供服务 [220](#bookmark263)

11.7.2 优化业务流程 [220](#bookmark264)

11.7.3 改善生活 [220](#bookmark265)

11.7.4 提高体育成绩 [221](#bookmark266)

11.7.5 优化机器和设备性能 [221](#bookmark267)

11.7.6 改善安全和执法 [221](#bookmark268)

11.7.7 金融交易 [221](#bookmark269)

11.7.8 电信业务 [221](#bookmark270)

11.7.9 销售 [222](#bookmark271)

11.8 大数据分析平台与工具 [222](#bookmark272)

11.8.1 大数据分析平台 [222](#bookmark273)

11.8.2 大数据分析的工具 [223](#bookmark274)

小结 [226](#bookmark275)

**第12章** **大数据挖掘** [227](#bookmark276)

12.1 数据挖掘概述 [227](#bookmark277)

12.1.1 数据挖掘的几个概念 228

12.1.2 数据挖掘对象与过程 [229](#bookmark278)

12.1.3 数据挖掘的常用方法 [232](#bookmark279)

12.1.4 数据挖掘的几个问题…… 234

12.1.5 数据挖掘的经典算法…… 237

12.2 大数据挖掘技术 [238](#bookmark280)

12.2.1 大数据挖掘关键技术 [239](#bookmark281)

12.2.2 大数据挖掘策略 [240](#bookmark282)

12.3 大数据挖掘应用 [243](#bookmark283)

12.3.1 市场营销 [243](#bookmark284)

12.3.2 销 售 [244](#bookmark285)

12.3.3 物流 [245](#bookmark286)

12.3.4 CRM [246](#bookmark287)



**目** **录**

小结 [248](#bookmark288)

**第** **13章** **大数据可视化** [249](#bookmark289)

13.1 可视化技术概述 [249](#bookmark290)

13.1.1 可视化技术的产生与发展 250

13.1.2 科学可视化 [253](#bookmark291)

13.1.3 信息可视化 [254](#bookmark292)

13.1.4 数据可视化 [257](#bookmark293)

13.2 大数据科学可视化 [260](#bookmark294)

13.2.1 高可伸缩性的分布式并行

可视化算法 [260](#bookmark295)

13.2.2 并行图像合成算法 [261](#bookmark296)

13.2.3 并行颗粒跟踪算法 [261](#bookmark297)

13.2.4 重要信息的提取与显示 262

13.2.5 原位可视化 [263](#bookmark298)

13.3 大数据可视化分析 [264](#bookmark299)

13.3.1 大数据可视化分析概念… 264

13.3.2 大数据可视化分析方法… 264

小结 [267](#bookmark300)

**第14章** **大数据安全** [268](#bookmark301)

14.1 大数据安全概述 [268](#bookmark302)

14.1.1 数据安全的定义 [269](#bookmark303)

14.1.2 数据处理与存储的安全 [269](#bookmark304)

14.1.3 数据安全的基本特点 [269](#bookmark305)

14.1.4 威胁数据安全的主要因素 270

14.1.5 安全制度与防护技术 [271](#bookmark306)

14.1.6 应用 [273](#bookmark307)

14.2 大数据安全的内容 [278](#bookmark308)

14.2.1 大数据的不安全因素 278

14.2.2 大数据安全的关键问题 279

14.2.3 大数据安全措施 [280](#bookmark309)

14.3 云安全 [281](#bookmark310)

14.3.1 云计算中用户的安全需求 281

14.3.2 威胁模型 [282](#bookmark311)

14.3.3 云安全的支撑技术 [282](#bookmark312)

14.3.4 用户数据隐私保护 [283](#bookmark313)

14.3.5 云计算执行环境的

可信性 [283](#bookmark314)



**大数据基础与应用**

14.3.6 资源共享问题 [284](#bookmark315)

小结 [284](#bookmark316)

**第15章** **大数据机器学习** [285](#bookmark317)

15.1 机器学习概述 [285](#bookmark318)

15.1.1 机器学习的产生与发展 [285](#bookmark319)

15.1.2 机器学习类型 [288](#bookmark320)

15.1.3 知识表示形式 [291](#bookmark321)

15.2 大数据机器学习的特点与

评测指标 [292](#bookmark322)

15.2.1 大数据机器学习的特点 293

15.2.2 大数据机器学习的

评测指标 [294](#bookmark323)

15.3 大数据机器学习的应用 [295](#bookmark324)

15.3.1 基于大数据的空气

质量推断 [295](#bookmark325)

15.3.2 人与建筑的关系分析 [296](#bookmark326)

15.3.3 针对全球问题的预测模型 296

15.3.4 全球地表覆盖制图可视化与

数据分析 [296](#bookmark327)

小结 [297](#bookmark328)

**第16章** **大数据推荐技术** [298](#bookmark329)

16.1 推荐技术概述 [298](#bookmark330)

16.1.1 推荐系统的产生与发展 299

16.1.2 推荐系统的概念 [299](#bookmark331)

16.1.3 推荐系统架构 [300](#bookmark332)

16.1.4 推荐系统类型 [300](#bookmark333)

16.1.5 推荐系统的评判标准 [302](#bookmark334)

16.2 推荐算法与推荐模式 [303](#bookmark335)

16.2.1 推荐算法 [303](#bookmark336)

16.2.2 推荐模式 [306](#bookmark337)

16.2.3 下一代推荐系统 [308](#bookmark338)

16.3 大数据推荐技术的挑战 [310](#bookmark339)

16.3.1 数据稀疏性问题 [310](#bookmark340)

16.3.2 大数据冷启动 [310](#bookmark341)

16.3.3 多样性与精确性的

两难命题 [310](#bookmark342)

16.3.4 大数据处理与增量

计算问题 [311](#bookmark343)

16.3.5 推荐系统的脆弱性问题 311

16.3.6 推荐系统效果评估 [311](#bookmark344)

16.3.7 用户行为模式的挖掘和

利用 [312](#bookmark345)

16.3.8 用户界面与用户体验 [312](#bookmark346)

16.3.9 多维数据的交叉利用 [313](#bookmark347)

16.3.10 社会推荐 [313](#bookmark348)

16.4 大数据人才推荐系统 [314](#bookmark349)

小结 [315](#bookmark350)

**第17章** **数据科学与数据思维** [**316**](#bookmark351)

17.1 数据科学概述 [316](#bookmark352)

17.1.1 数据科学的定义与信息化

过程 [316](#bookmark353)

17.1.2 数据科学的研究内容 [317](#bookmark354)

17.1.3 数据科学的研究过程与

体系框架 [318](#bookmark355)

17.2 大数据研究方式 [319](#bookmark356)

17.2.1 大数据分析的是全面的

数据 [320](#bookmark357)

17.2.2 重视数据的复杂性与弱化

精确性 [321](#bookmark358)

17.2.3 关注数据的相关性而

非因果关系 [321](#bookmark359)

17.3 数据专家 [322](#bookmark360)

17.3.1 数据科学家 [322](#bookmark361)

17.3.2 数据处理工程师 [325](#bookmark362)

17.3.3 大数据思维 [325](#bookmark363)

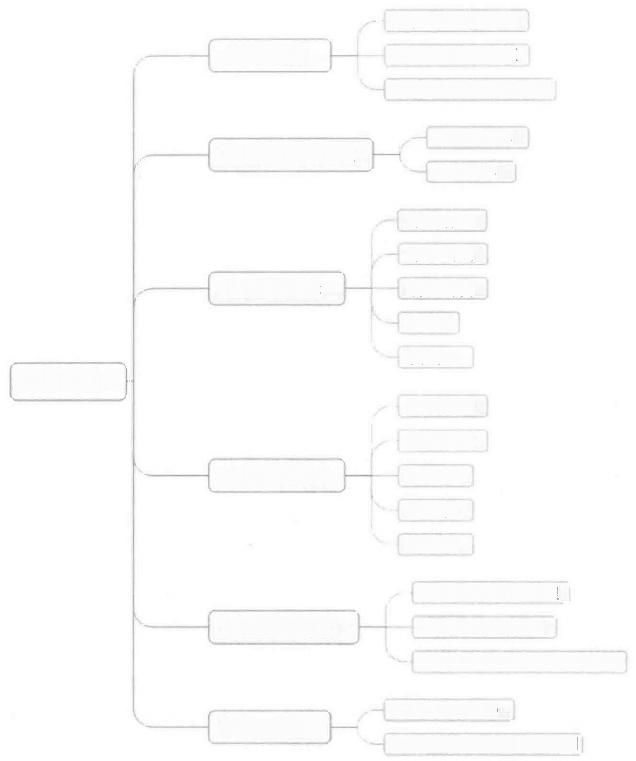
小结 [326](#bookmark364)

**参考文献** [327](#bookmark365)



**第1章** **走进大数据时代**

本章主要内容



电子数据迅速增加

数据中蕴含的价值

数据是国家的核心资产

互联网世界

大数据的生态环境

物理世界

数据容量

数据类型

价值密度

速度

真实性

走进大数据时代

非结构性

不完备性

时效性

安全性

可靠性

大数据处理的全过程

大数据技术的特征

大数据的关键问题与关键技术

大数据应用趋势

大数据应用

大数据应用评价与应用实例

大数据技术概述

大数据的性质

大数据的概念

应对大数据

需求是科学技术发展的原动力，大数据中蕴含着巨大而重要的价值，大数据的研究 已经成为计算机科学与技术研究的新热点，并显示出日益强大的生命力和吸引力。科学 大数据催生了数据密集型知识发现的第四科学研究范式的出现。大数据的出现，不仅带 来了机遇，也带来了困难和挑战。无论是科学界，还是企业界，都对大数据所带来的巨 大冲击寄予厚望，大数据是继20世纪末、21世纪初互联网蓬勃发展以来的又一轮新的 IT 工业革命。



大数据基础与应用

**1.1** **应对大数据**

在全球范围内，以电子方式存储的数据(简称为电子数据)总量空前巨大。2011年， 电子数据总量达到1.8ZB, 比2010年同期提高了1ZB, 统计结果表明，每经过两年就可以 增加一倍，预计到2020年可达到35ZB, 其预测曲线如图1-1所示。面对数据增长速度飞 快地提升，数据量狂增，对大量电子数据的高效存储、高效传输与快速处理成为必须解决 的问题。

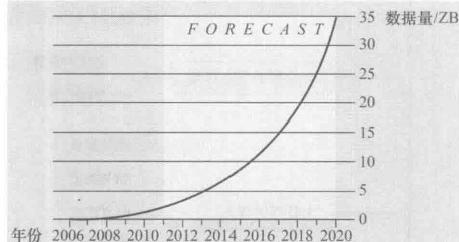


图1-1 全球数据总量的预测示意图

**1.1.1** **电子数据迅速增加**

云计算、物联网、移动互联网、手机、平板电脑和PC 机中的数据，各种公开的数据、 大型电子商务数据、遍布全球的各种传感器数据、工业蓬勃发展产生的工业数据、大型科 学研究设备产生的数据，以及社交媒体的快速发展，构成了大数据持续产生的基本生态环 境。尤其是近年来，随着互联网技术的发展，来自人们的日常生活，特别是来自互联网+X 服务的数据迅猛增加。据不完全统计，互联网当前包含93亿多个页面，80%～85%的数据 存储在数据库的文本中。互联网一天产生的全部内容可以刻满1.68 亿张 DVD, 发出的电 子邮件有2940亿封之多，发出的社区帖子达200万个，相当于《时代》杂志770年的文字 量。从统计角度来看，由于电子数据量迅速增加，预计中国数据技术和服务市场未来5年 的复合增长率将达51.4%,其中增长率最高的是存储市场，将达60.8%,服务器市场的增 长率则是38.3%,远远高于其他产品相关的市场。

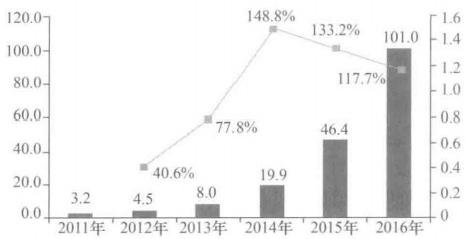
**1.1.2** **数据中蕴含的价值**

数据本身是无意义的，但是，有效地组织和使用数据，即对大数据进行交换、整合和分 析，可以发现新的知识、创造新的价值、推动科技大发展，凸显了数据是重要的战略资源。

据统计，2012年中国市场规模达到4.5亿元，2016年估计可达到上百亿元的规模， 如图1-2所示。



第 1 章 走进大数据时代



一市场规模(亿元)一同比增长率

图1-2 中国大数据应用市场规模与增长的预测

**1.1.3** **数据是国家的核心资产**

一个国家拥有数据的规模及分析获取价值的能力将成为综合国力的重要组成部分，存 储与控制的数据将成为国家陆权、海权和空权之外的另一个核心资产。

联合国在2012年发布了《大数据政务白皮书》,指出大数据对于联合国和各国政府来 说是一个历史性的机遇，通过使用丰富的数据资源，对社会经济进行前所未有的实时分析， 能够帮助政府更好地响应社会和经济运行。对数据的占有和控制也将成为国家之间和企业 之间的新的争夺点。社会需要大量大数据技术的专业人才，特别是数据分析复合型人才， 对这一专业人才的培养，将是一项十分重要的任务。

**1.2** **大数据的生态环境**

大数据是人类活动的产物，它来自人们改造客观世界的过程中，是生产与生活在网络 空间的投影。信息爆炸是对信息快速发展的一种逼真的描述，形容信息发展的速度如同爆 炸一般席卷整个空间。在20世纪40～50年代，信息爆炸主要指的是科学文献的快速增长。 而经过50年的发展，到20世纪90年代，由于计算机和通信技术的广泛应用，信息爆炸主 要指的是所有社会信息快速增长，包括正式交流过程和非正式交流过程所产生的电子式的 和非电子式的信息。而到21世纪的今天，信息爆炸是由于数据洪流的产生和发展所造成的。 在技术方面，新型的硬件与数据中心、分布式计算、云计算、高性能计算、大容量数据存 储与处理技术、社会化网络、移动终端设备、多样化的数据采集方式使大数据的产生和记 录成为可能。在用户方面，日益人性化的用户界面、信息行为模式等都容易作为数据量化 而被记录，用户既可以成为数据的制造者，又可以成为数据的使用者。可以看出，随着云 计算、物联网计算和移动计算的发展，世界上所产生的新数据，包括位置、状态、思考、 过程和行动等数据都能够汇入数据洪流，互联网的广泛应用，尤其是“互联网+”的出现， 促进了数据洪流的发展。

归纳起来，大数据主要来自互联网世界与物理世界。





**大数据基础与应用**

**1.2.1** **互联网世界**

大数据是计算机和互联网相结合的产物，计算机实现了数据的数字化，互联网实现了 数据的网络化，两者结合起来之后，赋予了大数据强大的生命力。随着互联网如同空气、 水、电一样无处不在地渗透人们的工作和生活，以及移动互联网、物联网、可穿戴联网设 备的普及，新的数据正在以指数级加速产生，目前世界上90%的数据是互联网出现之后迅 速产生的。来自互联网的网络大数据是指“人、机、物”三元世界在网络空间(Cyberspace) 中交互、融合所产生并可在互联网上获得的大数据，网络大数据的规模和复杂度的增长超 出了硬件能力增长的摩尔定律。

大数据来自人类社会，尤其是互联网的发展为数据的存储、传输与应用创造了基础与 环境。依据基于唯象假设的六度分隔理论而建立的社交网络服务 (Social Network Service, SNS), 以认识朋友的朋友为基础，扩展自己的人脉。基于Web 2.0 交互网站建立的社交网 络，用户既是网站信息的使用者，也是网站信息的制作者。社交网站记录人们之间的交互， 搜索引擎记录人们的搜索行为和搜索结果，电子商务网站记录人们购买商品的喜好，微博 网站记录人们所产生的即时的想法和意见，图片视频分享网站记录人们的视觉观察，百科 全书网站记录人们对抽象概念的认识，幻灯片分享网站记录人们的各种正式和非正式的演 讲发言，机构知识库和期刊记录学术研究成果等。归纳起来，来自互联网的数据可以划分 为下述几种类型。

**1.** **视频图像**

视频图像是大数据的主要来源之一，电影、电视节目可以产生大量的视频图像，各种 室内外的视频摄像头昼夜不停地产生巨量的视频图像。视频图像以每秒几十帧的速度连续 记录运动着的物体， 一个小时的标准清晰视频经过压缩后，所需的存储空间为GB 数量级， 对于高清晰度视频所需的存储空间就更大了。

**2.** **图片与照片**

图片与照片也是大数据的主要来源之一，截至2011年9月，用户向脸书(Facebook, 美 国的一个社会网络服务网站)上传了1400亿张以上的照片。如果拍摄者为了保存拍摄时的 原始文件，平均每张照片大小为1MB, 则这些照片的总数据量约为1.4×10¹²×1MB=140PB, 如果单台服务器磁盘容量为10TB, 则存储这些照片需要14000台服务器，而且这些上传的 照片仅仅是人们拍摄到的照片的很少一部分。此外，许多遥感系统24小时不停地拍摄并产 生大量照片。

**3.** **音频**

DVD 光盘采用了双声道16位采样，采样频率为44.1kHz, 可达到多媒体欣赏水平。如 果某音乐剧的时间为5.5min, 计算其占用的存储容量为：

存储容量=(采样频率×采样位数×声道数×时间)/8 =(44.1×1000×16×2×5.5×60)/8

≈55.5MB





第 1 章 走进大数据时代

**4.** **日志**

网络设备、系统及服务程序等，在运作时都会产生 log 的事件记录。每一行日志都记 载着日期、时间、使用者及动作等相关操作的描述。Windows 网络操作系统设有各种各样 的日志文件，如应用程序日志、安全日志、系统日志、Scheduler 服务日志、FTP 日志、 Www 日 志 、DNS 服务器日志等，这些根据系统开启的服务的不同而有所不同。用户在系 统上进行一些操作时，这些日志文件通常记录了用户操作的一些相关内容，这些内容对系 统安全工作人员相当有用。例如，有人对系统进行了IPC 探测，系统就会在安全日志里迅 速地记下探测者探测时所用的IP、时间、用户名等，用FTP 探测后，就会在 FTP日志中记 下IP、 时间、探测所用的用户名等。

网站日志记录了用户对网站的访问，电信日志记录了用户拨打和接听电话的信息，假 设有5亿用户，每个用户每天呼入呼出10次，每条日志占用400B, 并且需要保存5年， 则数据总量为5×10×365×400×5 Byte≈3.65PB。

**5.** **网页**

网页是构成网站的基本元素，是承载各种网站应用的平台。通俗地说，网站就是由网 页组成的，如果只有域名和虚拟主机而没有制作任何网页，客户仍旧无法访问网站。网页 要通过网页浏览器来阅读。文字与图片是构成一个网页的两个最基本的元素。可以简单地 理解为：文字就是网页的内容，图片就是网页的美观描述。除此之外，网页的元素还包括 动画、音乐、程序等。

网页分为静态网页和动态网页。静态网页的内容是预先确定的，并存储在 Web服务器 或者本地计算机、服务器之上，动态网页取决于用户提供的参数，并根据存储在数据库中 的网站上的数据而创建。通俗地讲，静态页是照片，每个人看都是一样的，而动态页则是 镜子，不同的人(不同的参数)看都不相同。

网页中的主要元素有感知信息、互动媒体和内部信息等。感知信息主要包括文本、图 像、动画、声音、视频、表格、导航栏、交互式表单等。互动媒体主要包括交互式文本、 互动插图、按钮、超链接等。内部信息主要包括注释，通过超链接链接到某文件、元数据 与语义的元信息，字符集信息，文件类型描述，样式信息和脚本等。

网页内容丰富，数据量巨大，每个网页有25KB 数据，则一万亿个网页的数据总量约 为25PB。

**1.2.2** **物理世界**

来自物理世界的大数据又被称为科学大数据，科学大数据主要来自大型国际实验：跨 实验室、单一实验室或个人观察实验所得到的科学实验数据或传感数据。最早提出大数据 概念的学科是天文学和基因学，这两个学科从诞生之日起就依赖于基于海量数据的分析方 法。由于科学实验是科技人员设计的，数据采集和数据处理也是事先设计的，所以不管是 检索还是模式识别，都有科学规律可循。例如希格斯粒子，又称为“上帝粒子”的寻找，



**大数据基础与应用**

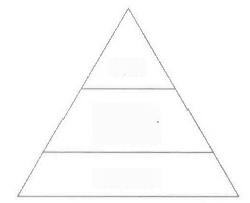
采用了大型强子对撞机实验。这是一个典型的基于大数据的科学实验，至少要在1万亿个 事例中才可能找出一个希格斯粒子。从这一实验可以看出，科学实验的大数据处理是整个 实验的一个预定步骤，这是一个有规律的设计，发现有价值的信息可在预料之中。大型强 子对撞机每秒生成的数据量约为1PB。建设中的下一代巨型射电望远镜阵每天生成的数据 量大约在1EB。波音发动机上的传感器每小时产生20TB 左右的数据量。

随着科研人员获取数据方法与手段的变化，科研活动产生的数据量激增，科学研究已 成为数据密集型活动。科研数据因其数据规模大、类型复杂多样、分析处理方法复杂等特 征，已成为大数据的一个典型代表。大数据所带来的新的科学研究方法反映了未来科学的

行为研究方式，数据密集型科学研究将成为科学研究的普 遍范式。

利用互联网可以将所有的科学大数据与文献联系在一 起，创建一个文献与数据能够交互操作的系统，即在线科 学数据系统，如图1-3所示。

对于在线科学数据，由于各个领域互相交叉，不可避 免地需要使用其他领域的数据。利用互联网能够将所有文 献与数据集成在一起，可以实现从文献计算到数据的整合。 这样可以提高科技信息的检索速度，进而大幅度地提高生 产力。也就是说，在线阅读某人的论文时，可以查看他们



文献

推导和组

合的数据

原始数据

图1-3 在线科学数据系统示意图

的原始数据，甚至可以重新分析，也可以在查看某些数据时查看所有关于这一数据的文献。

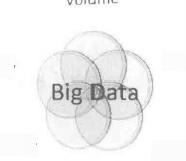
**1.3** **大数据的概念**

大数据是指数据规模大，尤其是指因为数据形式多样性、非结构化特征明显，导致数

据存储、处理和挖掘异常困难的那类数据集。大数据增 长快速，类型繁多，如文本、图像、视频等。大数据处 理包含数千万个文档、数百万张照片或者工程设计图的 数据集，如何快速访问数据成为核心挑战。无法用常规 的软件工具捕捉与处理。

通常将大数据归纳为5个“V”:Volume (数据容量)、 Variety(数据类型)、Viscosity (价值密度)、Velocity( 速 度)、Veracity (真实性),如图1-4所示。

**1.3.1** **数据容量**

Volume

Velocity

Veracity

Variety

Viscosity

图1-4

大数据的5个“V”

Volume 代表大数据的数据量巨大，存储容量单位的定义如表1-1所示。



第1章走进大数据时代

**表1-1** **存储容量单位的定义**

|  |  |  |  |
| --- | --- | --- | --- |
| 单 位 | 定 义 | 字节数(二进制) | 字节数(十进制) |
| Kilobyte(千) | 1024 Byte | 2l⁰ | 10³ |
| Megabyte(兆) | 1024 Kilobyte | 220 | 10⁶ |
| Gigabyte(吉) | 1024 Megabyte | 230 | 10° |
| Terabyte(太) | 1024 Gigabyte | 240 | 10¹² |
| Petabyte(拍) | 1024 Terabyte | 250 | 10¹⁵ |
| Exabyte(艾) | 1024 Petabyte | 260 | 10¹8 |
| Zettabyte(泽) | 1024 Exabyte | 270 | 10²l |
| Yottabyte(尧) | 1024 Zettabyte | 280 | 10²4 |

一般说来，超大规模数据是指在GB (即10°)级的数据，海量数据是指TB (即10²) 级的数据，而大数据则是指PB (即10¹⁵)级及其以上的数据。可以想象，随着存储设备容 量的增大，存储数据量的增多，大数据的容量指标是动态增加的，也就是说还会增大。下 一代计算机存储单位还会出现 Brontobyte 、Gegobyte 等存储单位。

2011年数据的总和是1.8ZB, 如果分别用9GB 的DVD 光盘和1TB 的2.5英寸硬盘来 保存，所需的光盘数量和硬盘数量如表1-2所示。

**表1-2** **存储1.8ZB 数据的比较**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **所** **用** **介** **质** | 单个容量(GB) | 所需数量(张) | 单个厚度(mm) | 堆叠高度(km) |
| DVD光盘 | 9 | 219902325555 | 1.2 | 263882.79 |
| 2.5英寸硬盘 | 1024 | 1932735283 | 9 | 17394.62 |

为了更形象地表示表1-2给出的结果，做说明如下：如果全部用9GB 的 DVD光盘来 保存，则所用的9GB的 DVD 光盘叠加后的高度超过2.6×10°km, 这个数字几乎是地球到 月球距离的三分之二。如果用1TB 的2.5英寸硬盘保存这1.8ZB 的数据，则所用1TB 的 2.5英寸硬盘叠加起来的高度超过1.7×10\*km, 几乎接近赤道周长的二分之一。为了进一步 说明，下面给出一个实际的数据。在计算机报上看到某银行的20 个数据中心有大约7PB 磁盘和超过20PB的磁带存储量，而且每年有50%～70%存储量的增长，存储27PB 数据需 要大约40万张80GB 的硬盘。

再如，如果1TB的硬盘的标准质量是670g, 那么存储1NB 的数据的硬盘总质量为：

1NB×0.67/10000=2°TB×0.67/10000=77245740809(万吨)

其中，1NB=1152921504606846976TB。

也就是说，存储1NB 的数据的硬盘，需要运载量为56万吨的巨型海轮最少来回运送 1379388229次，即约14亿次，估计当完成任务时，1000艘56万吨的巨型海轮都已损坏。

可以看出，上述例子中的数据是非常惊人的。如果用磁盘来存储大数据将是一项困难 的工作，因此不能用传统的方法来存储与管理大数据。





大数据基础与应用

**1.3.2** **数据类型**

Variety 代表大数据的数据类型繁多。由于大数据主要来自互联网，所以大数据包含多 种数据类型，如各种声音和电影文件、图像、文档、地理定位数据、网络日志、文本字符 串文件、元数据、网页、电子邮件、社交媒体供稿、表格数据等。其中，视频、图片和照 片日志为非结构化数据，网页为半结构化数据。

**1.3.3** **价值密度**

Viscosity 代表价值密度。 一般通过对大数据获取、存储、抽取、清洗、集成、挖掘与 分析来获得价值，大数据价值密度低，大概80%甚至90%的数据都是无效数据。以视频为 例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，难以进行预测分析、运营智 能、决策支持等计算，通常利用价值密度比来描述这一特点。随着物联网的广泛应用，信 息感知无处不在，信息海量，如何通过强大的计算机算法更迅速地完成数据的价值提纯， 是亟待解决的难题。

**1.3.4** **速度**

Velocity 代表大数据产生的速度快、变化的速度快。Facebook 每天产生25亿个以上条 目，每天增加数据超过500TB, 这样的变化率产生的数据需要快速处理，进而创造出价值。 传统技术不能够完成大数据高速存储、管理和使用。因此，应该研究新的方法与技术。如 果数据创建和聚合速度非常快，就必须使用迅速的方式来揭示其相关的模式和问题。发现 问题的速度越快，就越有利于从大数据分析中获得更多的机会与结果。

**1.3.5** **真实性**

Veracity 代表数据真实性。真实性是指数据是所标识的数据，而不是假冒的。准确性是 真实性的描述，不真实的数据需要清洗、集成和整合之后，获得高质量的数据，再进行分 析。也就是说，采集来的大数据不能保证完全真实，但是，大数据分析需要真实的数据。 越真实的数据，则数据质量越高。

**1.4** **大数据的性质**

从大数据的定义中可以看出，大数据具有规模大、种类多、速度快、价值密度低和真 实性差等特点，在数据增长、分布和处理等方面具有更多复杂的性质，如下所述。

**1.4.1** **非结构性**

结构化数据可以在结构数据库中存储与管理，并可用二维表来表达实现的数据。这类

**第1章** **走进大数据时代**

数据是先定义结构，然后才有数据。结构化数据在大数据中所占比例较小，只占15%左右， 现已应用广泛，当前的数据库系统以关系数据库系统为主导，例如银行财务系统、股票与 证券系统、信用卡系统等。

非结构化数据是指在获得数据之前无法预知其结构的数据，目前所获得的数据85%以 上是非结构化数据，而不再是纯粹的结构化数据。传统的系统无法对这些数据完成处理，

从应用角度来看，非结构化数据的计算是计算机科学的前沿。大数据的高度异构也导致抽 取语义信息的困难。如何将数据组织成合理的结构是大数据管理中的一个重要问题。大量 出现的各种数据本身是非结构化的或半结构化的数据，如图片、照片、日志和视频数据等 是非结构化数据，而网页等是半结构化数据。大数据大量存在于社交网络、互联网和电子 商务等领域。另一方面，也许有90%的数据来自开源数据，其余的被存储在数据库中。大 数据的不确定性表现在高维、多变和强随机性等方面。股票交易数据流是不确定性大数据 的一个典型例子。结构化数据、非结构化数据、半结构化数据的比较如表1-3所示。

**表1-3** **结构化数据、非结构化数据、半结构化数据的比较**

|  |  |  |  |
| --- | --- | --- | --- |
| **对** **比** **项** | **结构化数据** | **非结构化数据** | **半结构化数据** |
| 定义 | 具有数据结构描述 | 不方便用固定结构来表 | 处于结构化数据和非结构化数据之 |
| 信息的数据 | 现的数据 | 间的数据 |
| 结构与数据 的关系 | 先有结构，再有数据 | 只有数据，无结构 | 先有数据，再有结构 |
| 示例 | 各类表格 | 图形、图像、音频、视频 信息 | HTML文档，它一般是自描述的， 数据的内容与结构混在一起 |

大数据激生了大量研究问题。非结构化和半结构化数据的个体表现、 一般性特征和基 本原理尚不清晰，这些需要通过数学、经济学、社会学、计算机科学和管理科学在内的多 学科交叉研究。对于半结构化或非结构化数据，例如图像，需要研究如何将它转化成多维 数据表、面向对象的数据模型或者直接基于图像的数据模型。还应说明的是，大数据每一 种表示形式都仅呈现数据本身的一个侧面表现，并非其全貌。

由于现存的计算机科学与技术架构和路线，已经无法高效处理如此大的数据，如何将 这些大数据转化成一个结构化的格式是一项重大挑战，如何将数据组织成合理的结构也是 大数据管理中的一个重要问题。

**1.4.2** **不完备性**

数据的不完备性是指在大数据条件下所获取的数据常常包含一些不完整的信息和错 误，即脏数据。在数据分析阶段之前，需要进行抽取、清洗、集成，得到高质量的数据之 后，再进行挖掘和分析。

**1.4.3** **时效性**

数据规模越大，分析处理的时间就会越长，所以高速度进行大数据处理非常重要。如



**大数据基础与应用**

果设计一个专门处理固定大小数据量的数据系统，其处理速度可能会非常快，但并不能适 应大数据的要求。因为在许多情况下，用户要求立即得到数据的分析结果，需要在处理速 度与规模间折中考虑，并寻求新的方法。

**1.4.4** **安全性**

由于大数据高度依赖数据存储与共享，必须考虑寻找更好的方法来消除各种隐患与漏 洞，才能有效地管控安全风险。数据的隐私保护是大数据分析和处理的一个重要问题，对 个人数据使用不当，尤其是有一定关联的多组数据泄露，将导致用户的隐私泄露。因此， 大数据安全性问题是一个重要的研究方向。

**1.4.5** **可靠性**

通过数据清洗、去冗等技术来提取有价值数据，实现数据质量高效管理以及对数据的 安全访问和隐私保护已成为大数据可靠性的关键需求。因此，针对互联网大规模真实运行 数据的高效处理和持续服务需求，以及出现的数据异质异构、非结构乃至不可信特征，数 据的表示、处理和质量已经成为互联网环境中大数据管理和处理的重要问题。

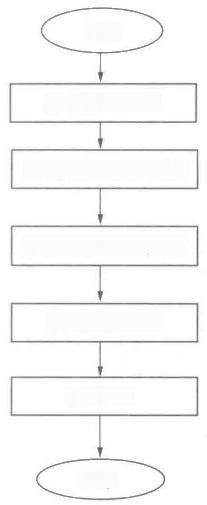
**1.5** **大数据技术概述**

大数据可分为大数据技术、大数据工程、大数据科学和大数据应用等领域。从解决问 题的角度出发，目前关注最多的是大数据技术和大数据应用。大数据工程是指大数据的规 划、建设运营和管理的系统工程，大数据科学关注大数据网络发展和运营过程中发现和验 证大数据的规律及其与自然和社会活动之间的关系。

大数据技术是指从数据采集、清洗、集成、分析与解释，进而从各种各样的巨量数据 中快速获得有价值信息的全部技术。目前所说的大数据有双重含义，它不仅指数据本身的 特点，也包括采集数据的工具、平台和数据分析系统。大数据研究的目的是发展大数据技 术并将其应用到相关领域，通过解决大数据处理问题来促进突破性发展。因此，大数据带 来的挑战不仅体现在如何处理大数据，并从中获取有价值的信息，也体现在如何加强大数 据技术研发，抢占时代发展的前沿。

被誉为数据仓库之父的比尔·恩门(Bill Inmon) 早在20世纪90年代就提出了大数据 的概念。近年来，互联网、云计算、移动计算和物联网迅猛发展，无所不在的移动设备、 RFID、无线传感器每分每秒都在产生数据，数以亿计用户的互联网服务时时刻刻在产生巨 量的交互，而业务需求和竞争压力对数据存储与管理的实时性、有效性又提出了更高要求。 在这种情况下，提出和应用了许多新技术，主要包括分布式缓存、分布式数据库、分布式 文件系统、各种 NoSQL 分布式存储方案等。





开始

数据采集与记录

数据抽取、清洗与标记

数据集成、转换与约简

数据分析与建模

数据解释

结束



第1章 走进大数据时代

**1.5.1** **大数据处理的全过程**

数据规模急剧扩大超过了当前计算机存储与处理能力。不

仅数据处理规模巨大，而且数据处理需求多样化。因此，数据

处理能力成为核心竞争力。数据处理需要将多学科结合，需要

研究新型数据处理的科学方法，以便在数据多样性和不确定性

的前提下进行数据规律和统计特征的研究。ETL 工具负责将分

布的异构数据源中的数据，如关系数据、平面数据文件等抽取

到临时中间层后进行清洗、集成、转换、约简，最后加载到数

据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。

一般说来，数据处理的过程可以概括为五个步骤，分别是

数据采集与记录，数据抽取、清洗与标记，数据集成、转换与

约简，数据分析与建模，数据解释，如图1-5所示。

**1.** **数据采集与记录**

数据的采集是指利用多个数据库来接收发自客户端

图1-5 大数据处理的全过程

(Web 、App 或者传感器形式等)的数据，并且用户可以通过

这些数据库来进行简单的查询和处理工作。例如，电子商务系

统使用传统的关系型数据库 MySQL 、SQL Server 和Oracle 等

结构化数据库来存储每一笔事务数据，除此之外， Redis 和

MongoDB 这样的 NoSQL 数据库也常用于数据的采集。在大数

据的采集过程中，其主要特点是并发率高，因为同时可能将有成千上万的用户来进行访问 和操作。例如，火车票售票网站和淘宝网站，它们并发的访问量在峰值时达到上百万，所 以需要在采集端部署大量数据库才能支撑，并且对这些数据库之间进行负载均衡和分片设 计。常用的数据采集方法如下所述。

(1)系统日志采集方法

很多互联网企业都有自己的海量数据采集工具，多用于系统日志采集，如 Hadoop 的 Chukwa 、Cloudera 的 Flume 、Facebook 的 Scribe 等，这些工具均采用分布式架构，能满足 每秒数百兆字节的日志数据采集和传输需求。

(2)网络数据采集方法

网络数据采集是指通过网络爬虫或网站公开API 等方式从网站上获取数据信息。该方 法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化 的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。

除了网络中包含的内容之外，对于网络流量的采集可以使用DPI 或 DFI 等带宽管理技 术进行处理。

(3)其他数据采集方法

对于企业生产经营数据或科学大数据等保密性要求较高的数据，可以通过与企业或研



**大数据基础与应用**

究机构合作，使用特定系统接口等相关方式采集数据。

**2.** **数据抽取、清洗与标记**

采集端本身设有很多数据库，如果要对这些数据进行有效的分析，应该将这些来自前 端的数据抽取到一个集中的大型分布式数据库，或者分布式存储集群，还可以在抽取基础 上做一些简单的清洗和预处理工作。也有一些用户在抽取时使用来自Twitter 的 Storm 对数 据进行流式计算，来满足部分业务的实时计算需求。大数据抽取、清洗与标记过程的主要 特点是抽取的数据量大，每秒钟的抽取数据量经常可达到百兆，甚至千兆数量级。

**3.** **数据集成、转换与约简**

数据集成技术的任务是将相互关联的分布式异构数据源集成到一起，使用户能够以透 明的方式访问这些数据源。在这里，集成是指维护数据源整体上的数据一致性，提高信息 共享利用的效率，透明方式是指用户不必关心如何对异构数据源进行访问，只关心用何种 方式访问何种数据即可。

**4.** **数据分析与建模**

统计与分析主要利用分布式数据库，或者分布式计算集群来对存储于其内的大数据进 行分析和分类汇总等，以满足大多数常见的分析需求。分析方法主要包括假设检验、显著性 检验、差异分析、相关分析、T 检验、方差分析、卡方分析、偏相关分析、距离分析、回归 分析(简单回归分析、多元回归分析)、逐步回归、回归预测与残差分析、曲线估计、因子 分析、聚类分析、主成分分析、判别分析、对应分析、多元对应分析(最优尺度分析)等。

在这方面， 一些实时性需求会用到 EMC 的 GreenPlum 、Oracle 的 Exadata 以及基于 MySQL 的列式存储Infobright 等，而一些批处理，或者基于半结构化数据的需求可以使用 Hadoop。统计与分析部分的主要特点是分析中涉及的数据量巨大，对系统资源，特别是/O 资源占用极大。

和统计与分析过程不同，数据挖掘一般没有预先设定好主题，主要是在现有数据上进 行基于各种算法的计算，起到预测的效果，从而实现一些高级别数据分析的需求，主要进 行分类、估计、预测、相关性分组或关联规则、聚类、描述和可视化、复杂数据类型挖掘 等。比较典型的算法有Kmeans 聚算法、SVM 统计学习算法和 NaiveBayes 分类算法，主要 使用的工具有 Hadoop 的 Mahout 等。该过程的特点主要是用于挖掘的算法很复杂，并且计 算涉及的数据量和计算量都很大，常用数据挖掘算法都以单线程为主。

建模的主要内容是构建预测模型、机器学习模型和建模仿真等。

**5.** **数据解释**

数据解释的目的是使用户理解分析的结果，通常包括检查所提出的假设并对分析结果 进行解释，采用可视化展现大数据分析结果。例如，利用云计算、标签云、关系图等呈现。

大数据处理的过程至少应该满足上述五个基本步骤，才能成为一个比较完整的大数据 处理过程。



第1章 走进大数据时代

**1.5.2** **大数据技术的特征**

大数据技术具有下述显著的特征。

**1.** **分析全面的数据而非随机抽样**

在大数据出现之前，由于缺乏获取全体样本的手段和可能性，针对小样本提出了随机 抽样的方法。在理论上，越随机抽取样本，就越能代表整体样本，但是获取随机样本的代 价极高，而且费时。出现数据仓库和云计算之后，获取足够大的样本数据，以至获取全体 数据成为可能并更为容易了。因为所有的数据都在数据仓库中，完全不需要以抽样的方式 调查这些数据。获取大数据本身并不是目的，能用小数据解决的问题绝不要故意增大数据 量。当年开普勒发现行星三大定律，牛顿发现力学三大定律都是基于小数据。从通过小数 据获取知识的案例中得到启发，人脑具有强大抽象能力，例如人脑就是小样本学习的典型。

2～3岁的小孩看少量图片就能正确区分马与狗、汽车与火车，似乎人类具有与生俱来 的知识抽象能力。从少量数据中如何高效抽取概念和知识是值得深入研究的方向。至少应 明白解决某类问题，多大的数据量是合适的，不要盲目追求超额的数据。数据无处不在， 但许多数据是重复的或者没有价值的，未来的任务不是获取越来越多的数据，而是数据的 去冗分类、去粗取精，从数据中挖掘知识、获得价值。

**2.** **重视数据的复杂性，弱化精确性**

对小数据而言，最基本和最重要的要求就是减少错误、保证质量。由于收集的数据少， 所以必须保证记录下来的数据尽量准确。例如，使用抽样的方法，就需要在具体的运算上 非常精确，在1亿人口中随机抽取1000人，如果在1000人的运算上出现错误，那么放大 到1亿人将会增大偏差，但在全体样本上，产生多少偏差就为多少偏差，不会被放大。

精确的计算是以时间消耗为代价的，在小数据情况下，追求精确是为了避免放大的偏 差而不得已为之。但在样本等于总体大数据的情况下，快速获得一个大概的轮廓和发展趋 势比严格的精确性重要得多。

大数据的简单算法比小数据更有效，大数据不再期待精确性，也无法实现精确性。

**3.** **关注数据的相关性，而非因果关系**

相关性表明变量 A 与变量 B 有关，或者说变量 A 的变化与变量 B 的变化之间存在一 定的比例关系，但在这里的相关性并不一定是因果关系。

亚马逊的推荐算法指出根据消费记录来告诉用户可能喜欢什么,这些消费记录有可能 是别人的，也有可能是该用户的历史购买记录，并不能说明喜欢的原因。不能说很多人都 喜欢购买A 和 B, 就存在购买A 之后的结果是购买B 的因果关系，这是一个未必的事情。 但其相关性高，或者说概率大。大数据技术只知道是什么,而不需知道为什么,就像亚马 逊的推荐算法指出的那样，知道喜欢A 的人很可能喜欢 B, 但却不知道其中的原因。知道 是什么就足够了，没有必要知道为什么。在大数据背景下，通过相互关系就可以比以前更 容易、更快捷、更清楚地进行分析，找到一个现象的关系物。系统相互依赖的是相互关系，





**大数据基础与应用**

而不是因果关系，相互关系可以表明将发生什么,而不是为什么发生，这正是这个系统的 价值。大数据的相互关系分析更准确、更快，而且不易受到偏见的影响。建立相互关系分 析法的预测是大数据的核心。当完成了相互关系分析之后，又不满足仅仅知道为什么,可 以再继续研究因果关系，找出原因。

**4.** **学习算法复杂度**

一般NlogN 、N²级的学习算法复杂度可以接受，但面对PB 级以上的海量数据，NlogN、 N²级的学习算法难以接受，处理大数据需要更简单的人工智能算法和新的问题求解方法。 普遍认为，大数据研究不只是上述几种方法的集成，应该具有不同于统计学和人工智能的 本质内涵。大数据研究是一种交叉科学研究，应体现其交叉学科的特点。

**1.5.3** **大数据的关键问题与关键技术**

**1.** **大数据的关键问题**

大数据来源非常丰富且数据类型多样，存储和分析挖掘的数据量庞大，对数据展现的 要求较高，并且重视处理大数据的高效性和可用性。

(1)非结构化和半结构化数据处理

如何处理非结构化和半结构化数据是一项重要的研究课题。如果把通过数据挖掘提取 粗糙知识的过程称为一次挖掘过程，那么将粗糙知识与被量化后的主观知识，包括具体的 经验、常识、本能、情境知识和用户偏好相结合而产生智能知识的过程就叫作二次挖掘。 从一次挖掘到二次挖掘是由量到质的飞跃。

由于大数据所具有的半结构化和非结构化特点，基于大数据的数据挖掘所产生的结构 化的粗糙知识(潜在模式)也伴有一些新的特征。这些结构化的粗糙知识可以被主观知识 加工处理并转化，生成半结构化和非结构化的智能知识。寻求智能知识反映了大数据研究 的核心价值。

(2)大数据复杂性与系统建模

大数据复杂性、不确定性特征描述的方法及大数据的系统建模这一问题的突破是实现 大数据知识发现的前提和关键。从长远角度来看，大数据的个体复杂性和随机性所带来的 挑战将促使大数据数学结构的形成，从而导致大数据统一理论的完备。从近期来看，应该 建立一种一般性的结构化数据和半结构化、非结构化数据之间的转化原则，以支持大数据 的交叉工业应用。管理科学，尤其是基于最优化的理论将在发展大数据知识的一般性方法 和规律性中发挥重要的作用。

现实世界中的大数据处理问题复杂多样，难以有一种单一的计算模式能涵盖所有不同 的大数据计算需求。研究和实际应用中发现，MapReduce 主要适合于进行大数据离线批处 理方式，不适应面向低延迟、具有复杂数据关系和复杂计算的大数据处理， Storm 平台适 合于在线流式大数据处理。

大数据的复杂形式导致许多与粗糙知识的度量和评估相关的研究问题。已知的最优化、 数据包络分析、期望理论、管理科学中的效用理论可以被应用到研究如何将主观知识融合





**第1章** **走进大数据时代**

到数据挖掘产生的粗糙知识的二次挖掘过程中，人机交互将起到至关重要的作用。 (3)大数据异构性与决策异构性影响知识发现

由于大数据本身的复杂性，致使传统的数据挖掘理论和技术已不适应大数据知识发现。 在大数据环境下，管理决策面临着两个异构性问题，即数据异构性和决策异构性问题。决 策结构的变化要求人们去探讨如何为支持更高层次的决策而去做二次挖掘。无论大数据带 来了何种数据异构性，大数据中的粗糙知识仍可被看作一次挖掘的范畴。通过寻找二次挖 掘产生的智能知识来作为数据异构性和决策异构性之间的连接桥梁。

寻找大数据的科学模式将带来对大数据研究的一般性方法的探究，如果能够找到将非 结构化、半结构化数据转化成结构化数据的方法，已知的数据挖掘方法将成为大数据挖掘 的工具。

**2.** **大数据的关键技术**

针对上述的大数据关键问题，大数据的关键技术主要包括流处理、并行化、摘要索引 和可视化。

(1)流处理

随着业务流程的复杂化，大数据趋势日益明显，流式数据处理技术已成为重要的处理 技术。应用流式数据处理技术可以完成实时处理，能够处理随时发生的数据流的架构。

例如，计算一组数据的平均值，可以使用传统的方法实现。对于移动数据平均值的计 算，不论是到达、增长还是一个又一个的单元，需要更高效的算法。但是想创建的是一个 数据流统计集，那需要对此逐步添加或移除数据块，进行移动平均计算。

(2)并行化

小数据的情形类似于桌面环境，磁盘存储能力在1GB～10GB 之间，中数据的数据量 在 1 0GB～1TB 之间，大数据分布式地存储在多台机器上，包含1TB 到多个 PB 的数据。 如果在分布式数据环境中工作，并且需要在很短的时间内处理数据，这就需要分布式处理。

(3)摘要索引

摘要索引是一个对数据创建预计算摘要，以加速查询运行的过程。摘要索引的问题是， 必须为要执行的查询做好计划。数据增长飞速，对摘要索引的要求远不会停止，不论是基 于长期还是短期考虑，必须对摘要索引的制定有一个确定的策略。

(4)可视化

数据可视化包括科学可视化和信息可视化。可视化工具是实现可视化的重要基础，可 视化工具包括两大类。

① 探索性可视化描述工具可以帮助决策者和分析师挖掘不同数据之间的联系，这是一 种可视化的洞察力。类似的工具有 Tableau 、TIBCO 和 QlikView 等。

② 叙事可视化工具可以独特的方式探索数据。例如，如果需要以可视化的方式在一个 时间序列中按照地域查看一个企业的销售业绩，可视化格式将被预先创建。数据将按照地 域逐月展示，并根据预定义的公式排序。





**大数据基础与应用**

**1.6** **大数据应用**

**1.6.1** **大数据应用趋势**

随着大数据技术逐渐应用于各个行业，基于行业的大数据分析应用需求也日益增长。 未来几年中针对特定行业和业务流程的分析应用将以预打包的形式出现，这将为大数据技 术供应商打开新的市场。这些分析应用内容还将覆盖很多行业的专业知识，也将吸引大量 行业软件开发公司的投入。

对于商业智能未来的趋势预测如图1-6所示。调查显示排在前三位的是丰富的挖掘模 型、实时的分析、精准的特定目的分析，其比例分别为27.22%、19.88%和19.11%。其后 是社交网络分析、云端服务和移动BI。

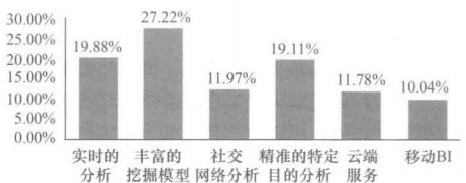


图1-6 商业智能发展的趋势

由以上趋势不难看出，在大数据时代，人们把焦点放在那些能快速改变现状的颠覆性 技术上，大数据存储与计算、数据挖掘与分析以及商业智能等应用前景远大。

**1.** **大数据细分市场**

大数据相关技术的发展，将创造出一些新的分市场。例如，以数据分析和处理为主的 高级数据服务，将出现以数据分析作为服务产品提交的分析即服务业务；将多种信息整合 管理，创造对大数据统一的访问和分析的组件产品；基于社交网络的社交大数据分析；将 出现大数据技能的培训市场，讲授数据分析课程，培养数据分析专门人才等。

**2.** **大数据推动企业发展**

大数据概念覆盖范围非常广，包括非结构化数据从存储、处理到应用的各个环节，与 大数据相关的软件企业也非常多，但是还没有哪一家企业可以覆盖大数据的各个方面。因 此，在未来几年中，大型IT 企业将为了完善自己的大数据产品线进行并购，首当其冲的将 是预测分析和数据展现企业等。

**3.** **大数据分析的新方法出现**

在大数据分析上，将出现新方法。就像计算机和互联网一样，大数据是新一波技术革



第 1 章 走进大数据时代

命。现有的很多算法和基础理论将产生新的突破与进展。

**4.** **大数据与云计算高度融合**

大数据处理离不开云计算技术，云计算为大数据提供弹性可扩展的基础设施支撑环境 以及数据服务的高效模式，大数据则为云计算提供了新的商业价值，大数据技术与云计算 技术必有更完美的结合。同样地，云计算、物联网、移动互联网等新兴计算形态，既是产 生大数据的地方，也是需要大数据分析方法的领域，大数据是云计算的延伸。

**5.** **大数据一体化设备陆续出现**

云计算和大数据出现之后，推出的软硬件一体化设备层出不穷。在未来几年里，数据仓 库一体机、NoSQL一体机以及其他一些将多种技术结合的一体化设备将进一步快速发展。

**6.** **大数据安全日益受到重视**

数据量的不断增加，对数据存储的物理安全性要求越来越高，从而对数据的多副本与 容错机制提出更高的要求。网络和数字化生活使得犯罪分子更容易获得关于人的信息，也 有了更多不易被追踪和防范的犯罪手段，可能会出现更高明的骗局。

**1.6.2** **大数据应用评价与应用实例**

大数据成功的应用将产生重大价值，需要研究判断大数据成功应用的标志。当前大 数据应用的研究关注国计民生的科学决策、应急管理(如疾病防治、灾害预测与控制、食 品安全与群体事件)、环境管理、社会计算以及知识经济等应用领域。

**1.** **判断大数据应用成功的指标**

(1)创造价值

大数据技术的应用应该能够创造切实的价值。据初步统计，大数据在医疗、政府、零 售以及制造产业上拥有上万亿的潜在价值。大数据应用的成功实现需要在附加收益、提升 客户满意度、削减成本等几个方面来考虑其带来的价值。因此，判断大数据应用成功的主 要指标是看其创造的价值。

(2)有本质提高

在模式上，大数据应用不仅是渐进式的商务模式改变，更重要的是在本质上的跳跃式 突破。例如，对初创企业来说，为了发现数据之间的关系，应用了机器学习算法使系统可 以进行调查， 一个社交推荐系统可以实时地给用户推荐有价值的位置信息，使用新的业务 模式去驱动位置信息类型业务。调查依赖大数据技术，同时从多于3000万个位置信息中获 取见解。现在的网站已经具备了理解人们之间如何进行互动的能力，并且位置信息也不只 局限于平台，而是真实世界。

(3)具备高速度

使用传统数据库技术会降低大数据技术的性能，同时也非常烦琐，因为不管这项技术 是否迎合使用者的需求，涉及的企业烦琐制度已远超想象。 一个成功的大数据应用，使用





**大数据基础与应用**

的工具集和数据库技术必须同时满足数据规模与多样性的双重需求。 一个 Hadoop 集群只 需几个小时就可以搭建，搭建完成后就可以提供快速的数据分析。事实上，大部分的大数 据技术都是开源的，这就表明可以根据需求添加支持和服务，同时许可完成快速部署。

(4)能完成以前所不能做的事情

在大数据技术出现之前，许多需求不可能实现，例如限时抢购。其原因是限时抢购网 站需要每日处理上千万用户的登录，将造成非常高的服务器负载峰值。通过高性能、快速 扩展的大数据技术可使这种商业模型成为可能。

综上所述，大数据应用成败的关键不是系统每秒可以处理多少数据量，而是应用大数 据之后创造了多少价值以及是否让业务有突破性的提升。专注业务类型，选择适合用户业 务的工具集才是应该重点关注的领域。

**2.** **大数据应用实例**

大数据技术应用广泛，几乎涉及各个领域。例如，网络大数据、金融大数据、健康医 疗大数据、企业大数据、政府管理大数据、安全大数据等，其趋势是从概念走向价值化的 大数据。大数据处理模式多样化并存，大数据安全隐私成为重要问题，大数据产业成为战 略性的新兴产业，数据商品化和数据共享联盟化的这种生态是未来一个重要的趋势，数据 科学兴起，大数据生态环境逐步发展。下面介绍大数据技术在不同的组织机构中的应用。

(1)医疗行业

① 医疗保健内容预测分析。利用医疗保健内容分析预测技术可以找到大量与病人相关 的临床医疗信息，通过大数据处理，能够更好地分析病人的信息。

② 早产婴儿的预测分析。在医院，针对早产婴儿，每秒钟有超过3000次的数据读取。 通过这些数据分析，医院能够提前知道哪些早产婴儿出现问题并且有针对性地采取措施， 避免早产婴儿夭折。

③ 精确诊断的预测分析。通过社交网络可以收集数据的健康类应用。也许未来数年后， 它们搜集的数据可使医生的诊断变得更为精确，例如不是通用的“成人每日三次、 一次一 片”,而是检测到人体血液中药剂已经代谢完成之后，自动提醒患者再次服药。

(2)能源行业

① 智能电网现在已经做到了终端，也就是所谓的智能电表。为了鼓励利用太阳能，在 家庭安装太阳能，除了卖电给用户，当家庭的太阳能有多余电的时候还可以买回来。通过 电网每隔5分钟或10分钟收集一次数据，收集来的数据可以用来预测客户的用电习惯等， 从而推断出在未来2～3个月时间内，整个电网大概需要多少电。有了这个预测后，就可以 向发电或者供电企业购买一定数量的电。因为电有点像期货，如果提前购买就会比较便宜， 购买现货就比较贵。通过这个预测，可以明显降低采购成本。

② 风力系统依靠大数据技术对气象数据进行分析，可以找出安装风力涡轮机和建设整 个风电场最佳的地点。以往需要数周的分析工作，现在利用大数据仅需要不足1小时便可 完成。

(3)通信行业

① 利用预测分析软件，可以预测客户的行为，发现行为趋势，并找出存在缺陷的环节，



**第** **1** **章** **走进大数据时代**

从而帮助公司及时采取措施，保留客户，可以减少客户流失率。此外，网络分析加速器通 过提供单个端到端网络、服务、客户分析视图的可扩展平台，帮助通信企业制定更科学合 理的决策。

② 电信业者透过数以千万计的客户资料，能分析出多种使用者的行为和趋势，卖给需 要的企业，这是全新的资料经济。

③ 通过大数据分析，对企业运营的全业务进行针对性的监控、预警、跟踪。系统在第 一时间自动捕捉市场变化，再以最快捷的方式推送给指定负责人，使他在最短时间内获知 市场行情。

④ 把手机位置信息和互联网上的信息结合起来，为顾客提供附近的餐饮店信息，接近 末班车时间时，提供末班车信息服务。

(4)交通行业

① 快递多效地利用了地理定位数据。为了使总部能在车辆出现晚点的时候跟踪到车辆 的位置和预防引擎故障，在货车上装有传感器、无线适配器和 GPS 。同时，这些设备也方 便公司监督管理员工并优化行车线路。为货车定制的最佳行车路径是根据过去的行车经验 总结而来的。

② 运输公司通过部署一系列的运输大数据应用，采集上千种数据类型，从油耗、胎压、 卡车引擎运行状况到GPS 信息等，甚至从司机们抱怨该系统的博客中收集数据，并通过分 析这些数据来优化车队管理、提高生产力、降低油耗，每年可节省大量的运营成本。

③ 车队通过汽车传感器在赛前的场地测试中实时采集数据，结合历史数据，通过预测 型分析发现赛车问题，并预先采取正确的赛车调校措施，降低事故概率并提高比赛胜率。

④ 缓解停车难问题。利用iOS 和 Android 手机，能够跟踪入网城市的停车位。用户只 需要输入地址或者在地图中选定地点，就能看到附近可用的车库或停车位、价格和时间区 间 。App 能够实时跟踪停车位的数量变化，能够实时监控多个城市的停车位。

⑤ 缓解道路拥堵的系统方案。基于实时交通报告来侦测和预测拥堵。当交管人员发现 某地即将发生交通拥堵，可以及时调整信号灯让车流以最高效率运行。这种技术对于突发 事件也很有用，例如帮助救护车尽快到达医院。而且随着运行时间的积累，这种技术还能 够学习过去的成功处置方案，并运用到未来预测中。

(5)零售业

大数据应用的必要条件在于IT 与经营的融合，范围可以小至一个零售门店的经营，大 至一个城市的经营。

① 收集社交信息，更深入地理解化妆品的营销模式，随后认识到必须保留两类有价值 的客户：高消费者和高影响者。希望客户通过接受免费化妆服务，进行口碑宣传，这是交 易数据与交互数据的完美结合，为业务挑战提供了解决方案。零售商用社交平台上的数据 充实了客户主数据，使其业务服务更具有目标性。

② 零售商也监控客户的店内走动情况以及与商品的互动，他们将这些数据与交易记录 相结合来展开分析，从而在销售哪些商品、如何摆放货品以及何时调整售价上给出意见， 此类方法已经帮助某零售企业减少了17%的存货，同时在保持市场份额的前提下，增加了 高利润率自有品牌商品的销售比例。



**大数据基础与应用**

③ 对零售商来说，孕妇是个含金量很高的顾客群体，通过顾客数据分析可以发现，怀 孕的妇女一般在怀孕第三个月的时候会购买很多无香乳液。几个月后，她们会购买镁、钙、 锌等营养补充剂。根据数据分析所提供的模型，可以制订全新的广告营销方案，在孕期的 每个阶段给客户寄送相应的优惠券。结果，孕期用品销售呈现了爆炸性的增长。大数据的 巨大威力引起了巨大轰动。

④ 在淘宝上每天进行数以万计的交易，相应的交易时间、商品价格、购买数量被记录。 更重要的是，这些信息可以与买方和卖方的年龄、性别、地址甚至兴趣爱好等个人特征信 息相联系。各大中小城市的百货大楼、商场做不到这一点，大大小小的超市做不到这一点， 而互联网时代的淘宝网可以做到。淘宝数据魔方就是淘宝平台上的大数据应用方案。通过 这一服务，商家可以了解淘宝平台上的行业宏观情况、自己品牌的市场状况、消费者行为 情况等，并可以据此进行生产、库存决策，而与此同时，更多的消费者也能以更优惠的价 格买到更心仪的商品。

⑤ 通过大数据分析发现，如果一个人在下午4点左右给汽车加油的话，他很可能在接 下来的1个小时内要去购物或者吃饭，而这1个小时的花费为150～200 元。商家正需要这 样的信息，因为这样他们就能在这个时间段的加油小票背面附上加油站附近商店的优惠券。

(6)金融

通过掌握的企业交易数据，借助大数据技术自动分析，判定是否给予企业贷款，全程 不出现人工干预。

资本市场公司每天的工作之一就是利用计算机程序分析全球3.4 亿微博账户的留言， 进而判断民众情绪，再进行打分。根据打分结果，决定如何处理手中数以千万元计的股票。 判断原则很简单，如果所有人似乎都高兴，那就买入；如果大家的焦虑情绪上升，那就抛售。

像VISA 这样的信用卡发行商，站在了信息价值链最佳位置上。VISA 的数据部门收集 和分析了来自210个国家的15亿信用卡用户的650亿条交易记录，用来预测商业发展和客 户的消费趋势，然后卖给其他公司。

**3.** **大数据** **CRM**

CRM 是企业的一项商业策略，它按照客户细分情况有效地组织企业资源，培养以客户 为中心的经营行为以及实施以客户为中心的业务流程，并以此为手段来提高企业的获利能 力、收入以及客户满意度。CRM 实现的是基于客户细分的一对一营销，所以是按照客户细 分原则对企业资源的有效组织与调配。以客户为中心是企业的经营行为和业务流程都要围 绕客户，通过CRM 手段来提高利润和客户满意度。

(1)CRM 的主要功能指标

① 客户概况分析：包括客户的层次、风险、爱好和习惯等。

② 客户忠诚度分析：指客户对某个产品或商业机构的信用程度、持久性和变动情况等。 ③ 客户利润分析：指不同客户所消费的产品的边缘利润、总利润额和净利润等。

④ 客户性能分析：指不同客户所消费的产品按种类、渠道、销售地点等指标划分的销 售额。

⑤ 客户未来分析：包括客户数量、类别等情况的未来发展趋势、争取客户的手段等。

第 1 章 走进大数据时代

⑥ 客户产品分析：包括产品设计、关联性、供应链等。

⑦ 客户促销分析：包括广告、宣传等促销活动的管理。

(2)CRM 与大数据融合

应用大数据技术可以从各种类型的数据中快速获取有价值的信息。CRM 作为客户关系 管理系统专家，可以应用大数据帮助企业获得客户资源的有效管理。

① CRM 将带动大数据市场快速成长。

大数据应用将进入传统行业，而CRM 将带动商业分析应用市场的快速成长。按照 CRM 的经营理念，企业应制定CRM 战略，进行业务流程再造，才能据以实施CRM 技术和应用 系统，从而增强客户满意度，培育忠诚客户，达到实现企业经营效益最大化的目标。在企 业的日常工作中， 一般的客户关系管理至少要涵盖营销管理、销售管理、客户服务和技术 支持四个层面的功能，以保证企业能适时与客户密切交流，处理好人、流程、技术三者的 关系。因此，客户关系管理系统不仅是一个管理理念的实现，更是一套人机交互系统和解 决方案，其中贯穿着系统管理、企业战略、人际关系合理利用等思想，它能帮助企业更好 地吸引潜在客户和留住最有价值的客户。通过在线CRM, 企业可以迅速发现客户，并有效 地维系客户，实现最大利益。

② 把握行业趋势，抢占市场先机。

随着数据源指数级增长，信息的数量及复杂程度快速扩大，从海量数据中提取信息的 能力正快速成为战略性的强制要求。可以看出，由于数据的爆发式增长，企业能够从这些 繁乱的数据中快速获得战略决策信息是战胜对手的关键。面对不断发展的数据，大数据的 挖掘和分析尤为重要。

中国有6.7亿的互联网用户、多样化的1.8万亿GB 数据，企业数据每年以55%的速度 增长。在蓬勃发展的中国市场环境中，大数据所带来的机遇前所未有，这将是中国市场的 营销者们预期取得回报的最佳时机。这也正是以数据为本，分析为先的CRM 发展的良机。

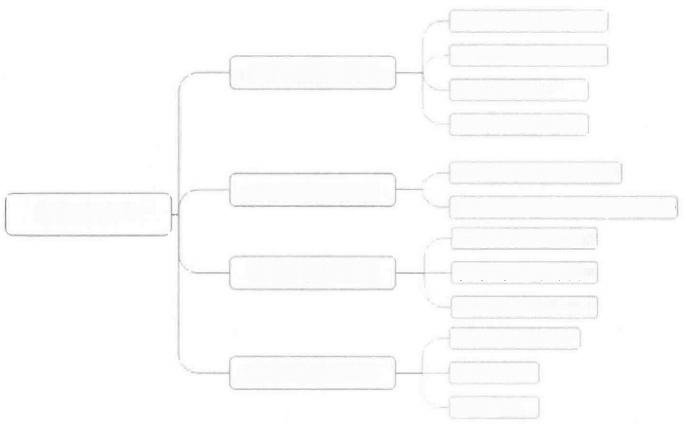
小结

本章主要概括性介绍了大数据的概念、大数据的主要特征、大数据的生态环境、大数 据技术、重要的大数据技术问题和大数据应用等。通过对本章内容的学习，可以为进一步 学习大数据技术建立初步的基础。



**第2章** **科学研究范式**

本章主要内容



科学实验特点与步骤

科学实验构成与分类

科学研究第一范式

科学实验程序

科学实验使用原则

科学理论的特征与价值

科学研究第二范式

科学理论的结构与体系建立方法

系统模拟发展过程

系统模拟基本方法

系统模拟语言

数据密集型计算

格雷法则

核心内容

科学研究第三范式

科学研究第四范式

科学研究范式

科学问题是指一定时代的科学认识主体，在已完成的科学知识与科学实践的基础之上， 提出的有可能解决的问题，包括求解目标和应答领域。科学发展的历史就是一个不断提出 科学问题和不断解决科学问题的历史。科学问题涵盖技术问题，技术问题是科学问题的子 集，科学问题具有时代性、混沌性、可解决性、可变异性和可待解性等特征，而科学问题 的方法论具有裂变作用、聚变作用与激励作用。研究科学问题的方法论异常重要。

万物之灵的人类对外部世界的认识已达到令人惊叹的高度，在宏观上远及亿万光年的 宇宙，在微观上已达层子、夸克世界。从宏观到微观、从自然到社会的观察、感知、计算、 仿真、模拟、传播等活动，产生出大数据。科学家不仅通过对广泛的数据实时、动态地监 测与分析来解决难以解决或不可触及的科学问题，更是把数据作为科学研究的对象和工具， 基于数据来思考、设计和实施科学研究。数据不再仅仅是科学研究的结果，而且变成科学 研究的活动基础。研究者不仅关心数据建模、描述、组织、保存、访问、分析、复用和建 立科学数据基础设施，更关心如何利用泛在网络及其内在的交互性、开放性，利用大数据的 可知识对象化、可计算化，构造基于数据的、开放协同的研究与创新模式，进而诞生了数据 密集型的知识发现的科学研究第四范式。数据科学家也就成为第四范式的实际践行者。

科学范式是科学发现运作的理论基础和实践的规范，是科学工作者共同遵循的普适的 世界观和行为方式。范式代表了人类思维的方式和根基，也是科学知识交流时共同遵守的 法则。范式的本质是理论体系，范式是一种公认的模型或模式。范式的演变是科学研究的

**第** **2** **章** **科学研究范式**

方法及观念的替代过程，科学的发展不是靠知识的积累而是靠范式的转换来完成，新范式 的形成表明建立起了常规科学。库恩的模型描述了一种科学的图景： 一组观念成为特定科 学领域的主流和共识，创造了一种关于这个领域的观念，进而拥有了自我发展的动力和对 这个领域发展的控制力。它代表了对观察到的现象的合理解释，这种观念或范式从渐进发 展的机制中获得启发和动力，同时被科学家逐渐完善。当现有范式无法解释观察到的现象， 或者实验最终证明范式是错误时，那么范式失败，转变范式的机会也就随之到来。大数据 的出现是科学研究第四范式出现的导火线。存储、处理、分析大数据的能力是科学必须适 应的新事实，数据是这个新范式的核心，它与实验、理论、模拟共同成为现代科学方法的 统一体。在科学发展的历史长河中，人类先后经历了实验、理论和计算三种科学研究范式， 这三种范式对科学与技术的发展作出了巨大的贡献，并已成功地将科学的发展引领至今天 的辉煌，而且仍处于现代科学的核心。毫无疑问，基于现有的范式与技术，科学研究还将 获得增量进展，传统的三种科学发现模式已经不能在一些领域进一步发挥有效的作用。如 果需要更重大的突破，就需要新的方法，需要开创新范式，科学研究第四范式应运而生。

大数据科学将给科学家带来技术挑战，IT 技术和计算机科学将在推动未来科学发现中 发挥重要作用。

**2.1** **科学研究第一范式**

科学实验主要描述自然现象，以观察和实验为依据的研究，可称之为经验范式。科学 实验是人们为实现预定目的，在人工控制条件下，通过干预和控制科研对象而观察和探索 科研对象的规律和机制的一种研究方法。科学实验是观察的一种形式。

近年来，科学实验经历了很大发展，科学实验的规模愈来愈大，科学实验再也不是科 学家个人的事业，而成为整个社会事业的一部分。科学实验不仅是搜集科学事实、获得感 性材料的基本方法，也是检验科学假说、形成科学理论的实践基础，二者互为补充。

**2.1.1** **科学实验特点与步骤**

**1.** **科学实验特点**

(1)纯化观察对象条件

可以利用各种实验手段对研究对象进行各种人工改变和人工控制，致使被研究对象的 特性暴露出来，就能获得被研究对象在自然状态下难以观察到的特性。

(2)强化观察对象

可以利用各种实验手段创造出在地球表面的自然状态下无法出现的或几乎无法出现的 特殊条件。在这种强化了的特殊条件下，发现了许多具有重大意义的新事实。

(3)可重复性

可以通过一定实验手段使被观察对象重复出现，这样有利于长期、反复比较观察，对 以往的实验结果加以核对。





**大数据基础与应用**

**2.** **科学实验步骤**

科学实验的方法应该包括如下步骤。

● 观察：观察事实和事件的详细记录。

● 定义：对问题进行定义，并且定义可操作的范围、手段。

● 假设：假设是指对一种事物或一种关系的暂时性解释。

● 检验：收集证据和检验假设， 一方面要能提供假设所需的客观条件， 一方面要找 到方法来测量相关参数。

● 发表：发表研究结果，科学信息必须公开，真正的科学关注的是解决问题。

● 建构：构建理论。孤立的问题无法建立理论，科学的理论是可以被证实的。

**2.1.2** **科学实验构成与分类**

**1.** **科学实验构成**

(1)实验者

实验者是指组织、设计和进行科学实验的人，主要负责实验目的的确定，实验方案的 设计、实验步骤的制定、实验过程的操作、实验结果的处理与解释等。实验者是实验活动 的主体，任何一个环节都不可脱离实验者。

(2)实验对象

实验对象可以是自然界的物体及其现象，也可以是人造的物体及其现象，实验对象处 于认识客体的地位。

(3)实验手段

实验手段是由实验的仪器、工具、设备等客观物质条件组成，实验仪器是其中的主要 成分。

**2.** **科学实验分类**

实验类型可以基于下述几点分类。

(1)基于实验目的分类

根据实验的目的，可以将科学实验分为定性实验、定量实验和结构分析实验。 (2)基于手段的作用分类

根据实验手段(仪器、设备工具等)是否直接作用于被研究对象，实验可分为直接实 验、间接实验和模型实验等。

(3)基于对象性质分类

根据实验对象性质的多样性，可以分为物理实验、化学实验、生命实验、人体实验等。 (4)基于预定目的分类

根据实验者的预定目的可分为定性实验、定量实验、测量实验、对照实验、验证性实 验、判定性实验和中间实验等。

(5)基于实验对象分类

根据实验对象的透明度，可以分为黑箱实验、灰箱实验和白箱实验等。







第2章 科学研究范式

**2.1.3** **科学实验程序**

**1.** **准备阶段**

科学实验过程的第一个阶段是准备阶段。准备阶段决定了科学实验的价值与成败。

● 确立实验目的；

● 着手实验设计；

● 实验仪器、设备与材料的准备。

**2.** **实施阶段**

实施阶段是实验者操作一定的仪器设备使其作用于实验对

象，以取得某种实验效果和数据。

**3.** **结果处理阶段**

科学实验的第三个阶段是结果处理阶段。在这一阶段，需要

对实验结果进行分析。

实验法是自然科学研究领域最早广泛应用的研究方法之一，

是近代自然科学建立的基础。达·芬奇、伽利略、牛顿(见图2-1)

等都充分地利用实验方法做出了巨大的科学成就。

图2-1 著名科学家牛顿

**2.1.4** **科学实验使用原则**

**1.** **掌握理论**

在设计实验方案和进行具体实验的过程中，离不开理论的指导和前人经验的积累。应 熟练掌握与实验课题有关的理论和经验。

**2.** **提出假设**

实验在科学研究中主要有两种目的： 一种目的是探索和发现新现象或新规律，另一种 目的是检验已有知识或理论的正确性。应事先提出假说或需要检验的观点与理论等。

**3.** **精心设计**

根据一定的理论，结合具体的研究对象，可以采取不同的研究方式。应精心设计，严 密组织。

**4.** **做好准备**

实验环境对于实验的成功与否有很大关系，应选择好实验环境，准备好实验工具。

**5.** **保持状态**

无论研究对象是自然界中的事物，还是人类自己，为了保持实验结果的客观性，要尽





**大数据基础与应用**

量保持受验者的常规状态。只有在常态下，事物或人所表现出来的才是其真实的情况。

**6.** **控制因素**

在实验过程中，要根据研究目的来尽量控制实验中的各种因素。要突出主要因素，排 除次要因素、偶然因素以及外界的干扰，从而能更准确地认识事物的本质规律。

**7.** **仔细观察**

应仔细观察，尽可能得到精确的数据。

**8.** **反复实验**

在做深入的大规模的实验前，首先要做探索性的实验，先简单后复杂，这样可以为以 后的实验工作积累相关的信息和思路。实验要注意其可重复性。只有多次重复，才能表明 其成果可以被大家认可。

**9.** **核对结论**

实验结束后，要对实验中获得的数据做进一步的处理与整理，从中提取出科学事实或 某种规律性的理论。在分析过程中，可以利用统计分析等方法，借助于计算工具等手段， 从数据之间的因果关系、起源关系、功能关系、结构关系等多角度、多层次、多因素地进 行处理。

**2.2** **科学研究第二范式**

科学研究第二范式是以建模和归纳为基础的理论学科和分 析范式，又称之为理论范式。科学理论是对某种经验现象或事实 的科学解说和系统解释，它是由一系列特定的概念、原理、命题 以及对它们的严密论证组成的知识体系。理论物理学家、思想家 及哲学家，也是相对论的创立者，阿尔伯特·爱因斯坦经常被认 为是现代物理之父及20世纪最重要的科学家之一，如图2-2所示。

经验现象是可以感觉到它，或通过观测、计量等获得有关它 的信息证明，这些信息也能被其他观察者所证实。经验现象具有 两个基本功能。

① 观察结果产生了解释这些事实的需要，即建立理论的 需要。

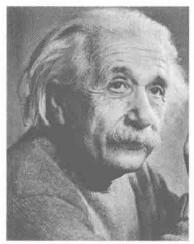


图2-2 著名理论物理学家

爱因斯坦

② 有助于对现存理论重新予以系统阐述、扩展和澄清。当观察到的现象与现存理论的 解释不完全一致时，就必须正视这种差异。

理论主要是解释事实，而不是描述事实。而且这种解释是对个别的、特殊的事件所做 的一般的、概括的总结性归纳论述，这就可以通过一般性的定理、规则来认识个别事物。 应说明的是，在真理面前，任何科学理论都是试探性的、暂时的、猜测性的，而且是试探





第2章科学研究范式

性的假说或真理的近似。

**2.2.1** **科学理论的特征与价值**

**1.** **科学理论的特征**

(1)抽象性

抽象性是科学理论的重要特征，是对经验事实的简化与概括。因为任何事实或现象都 受到多种因素的影响，在分析事实时，借助于理性思维的抽象对事实进行必要的简化，而 使事实易于分析，并以纯粹的形态呈现出来。

(2)逻辑性

科学理论建立在明确的概念、恰当的判断、正确的基于因果关系的推理与严密的逻辑 证明基础之上。

(3)系统性

由于科学理论是一种系统化的逻辑体系，系统性也是科学理论的重要特征。由于科学 理论建立在对经验事实的简化即一定的前提假设之上，所以任何科学理论都有其特定的边 界或有效范围。

**2.** **科学理论的价值**

科学理论具有解释和预测功能，科学理论的基本价值就在于提供关于事实的解释，引 导对事实的认识。

**2.2.2** **科学理论的结构与体系建立方法**

**1.** **科学理论的结构**

科学理论的结构由三个要素组成：概念、联系和原理。

(1)概念

概念是人们在对经验事实获得丰富而充分的感性认识基础上经反复抽象思维而形成的 反映事物本质联系的逻辑形式，是经验事实在人们头脑中的重构，具有抽象特征。建立概 念是理论建构必要的而且是最重要的环节。

(2)联系

一个科学理论所使用的概念往往不是单一的，而是一个甚至若干个概念群。这些概念 按照固有的隶属关系、包含关系、并列关系、联结关系等，形成一个有序的概念网络或概 念体系，并构成理论论述体系的关节点。

(3)原理

原理是指以概念为基础，对事实或现象进行分类和分析，概括或假设它们之间的逻辑 关系，并能给予合理的解释。原理是对经验事实基本关系的反映，是一种表现为科学判断 的思维形式。





**大数据基础与应用**

**2.** **科学理论体系建立方法**

(1)从抽象上升到具体的方法

从抽象上升到具体的方法是将科学研究已经获得的结果(概念、原理、规律等),按照 从低级到高级、从简单到复杂、从抽象到具体的上升过程加以系统化，进而构造一个严密 的科学理论体系的方法。

用公理化方法构建科学理论体系是该方法在科学认识中的重要作用之一，从尽可能少 的基本概念、公理、公设出发，运用演绎推理规则，推导出一系列的命题和定理，并依次 排列建立整个理论体系的方法。

(2)逻辑与历史相统一的方法

无论自然本身的发展过程还是人对自然认识的发展过程都按照从低级到高级、从简单 到复杂的方向运动，表明了逻辑与历史的统一，即从简单上升到复杂这个抽象思维的进程 符合现实的历史过程。 一个成熟的科学理论体系，不仅要实现从抽象到具体的上升，同时 也应达到历史的和逻辑的统一。

**2.3** **科学研究第三范式**

计算机的诞生催生了第三范式出现，阿兰·麦席森·图灵 (Alan Mathison Turing, 1912—1954)是英国数学家、逻辑学家，他被视为计算机之父，如图2-3所示。

科学研究第三范式是以模拟复杂现象为基础的计算科学范式，又称为模拟范式。程序 模拟是一种能用来帮助用户在不确定条件下进行决策的方法。用户必须在不完全了解事件 的发生及其影响如何的情况下，从若干方案中选出一种行动方案来。如果出现特殊事件， 将会有什么结果，这也有不确定性。

模拟有三种方法，即数学模拟方法、程序模拟方法和物理模拟方法。其中，数学模拟 方法比较抽象，物理模拟方法的代价较高，程序模拟方法处在两者之间。数学模拟方法与 程序模拟方法是常采用的方法。数学模拟方法是指以数学为工具的科学研究方法，用数学 语言表达事物的状态、关系和过程，并经过推导形成解释和判断。这种方法的基本特征是

高度抽象、高度精确，具有普遍意义。数学语言是科学研究的 简洁、精确、形式化语言。

程序模拟是系统科学方法中的主要方法，系统科学方法是 将研究的对象看成一个整体，并且使思维对应于适当的抽象级 别， 一般遵循整体性、动态性、最优化和模型化原则。常用的 系统科学方法有系统分析法、黑箱法、功能模拟法、整体优化 法和信息分析法等。

程序模似是建立研究对象的数学模型或描述模型并在计算 机上加以体现和试验。研究对象包括各种类型的系统，它们的 模型是指借助有关概念、变量、规则、逻辑关系、数学表达式、





图2-3 著名计算机科学家 阿兰·麦席森·图灵



第2章科学研究范式

图形和表格等对系统的一般描述。把这种数学模型或描述模型转换成对应的计算机上可执 行的程序，给出系统参数、初始状态和环境条件等输入数据后，可在计算机上进行运算得 出结果，并提供各种直观形式的输出，还可根据对结果的分析改变有关参数或系统模型的 部分结构，重新进行运算。

**2.3.1** **系统模拟发展过程**

当设计和构造复杂的系统时，或者研究自然界、人类社会中漫长的演变过程和不易重 复试验的事物时，如果对研究对象本身进行试验，从时间、人力、物力等因素考虑要付出 昂贵的代价，甚至不可能进行。因此，需要制造一个模型来进行各种试验。

为了对系统模拟，首先要确定或表达所要研究的系统。用数学模型能较方便地确定一 个系统，全面地反映对系统的已有认识或需要验证的假设，但缺乏直观性，也不便于进行 试验。在数学模型的基础上，可进一步做出实物模型，它体现人们所要求的真实系统有关 的性质，但在形式和规模上不必与真实系统完全一致。用实物模型试验比较直观、可信， 但仍不够经济和方便。

计算机出现以后，可以把数学模型编制成计算机程序，提供新的、通用的试验方法。

计算机也可用于模拟与运筹有关的活动。它的应用领域很快就扩展到各种类型的系统，从 规模巨大的系统一直到小型的系统，这些系统的数学描述常常非常复杂，要给出完全的解 析解或精确的数值解非常困难。程序模拟通过反复试验，帮助人们了解系统的性能，检验 预想的假设，进行系统分析、设计、预测或评估，还可提供相当逼真的环境，借以培养和 训练人员。程序模拟已成为工程研制、自然学研究、经济和社会问题研究、教学训练活动、 军事研究、组织管理等许多领域中的一个有力的工具。著名的四色问题就是利用程序模拟 方法成功解决的。

**2.3.2** **系统模拟基本方法**

程序模拟一般从形成问题到最后模型确认的步骤如下。

① 形成问题，明确模拟的目的和要求。

② 数据收集，尽可能收集和处理与系统有关的数据。

③ 形成数学模型，找出组成系统的各个部件，并描述与各时刻的状态有关的变量或参数。 ④ 根据收集的数据确定或估计模型中的参数，并选择模型的初始状态。

⑤ 设计逻辑或信息的流程图，直至编制出计算机程序。

⑥ 程序验证，检验程序与数学模型之间的一致性，以及输入量的合理性。

⑦ 进行模拟试验，对给定的输入在计算机上执行程序。

⑧ 结果分析，收集和整理试验结果并做出解释。

⑨ 模型确认，检验由模型所得的结果与真实系统的性能数据的一致性程度。 模型有效的级别可分为：

● 重现有效的，即模型可重现真实系统的性能；



大数据基础与应用

● 预测有效的，即模型能有效地预测真实系统的未来性能；

● 构成有效的，即模型能反映真实系统内部的结构。

由于系统本身是随时间变化的，对真实系统数据和模型试验结果的比较常需要采用时 间序列分析方法或统计分析的方法。

**2.3.3** **系统模拟语言**

模拟程序可采用汇编语言、通用编程语言(如 Java 、C 、C++等),还可采用专用模拟 语言编写。程序模拟语言是一种高级的描述系统模型的编程语言，可以提供表示系统模型 中许多基本单元、部件和调度操作的模块。用户可较方便地确定模型的基本结构，只要再 添加一些辅助的程序就可设计出模拟程序。

模拟语言一般是在别的通用编程语言的基础上建立的，需要编译程序进行预编译，把模 拟语言程序转换成通用编程语言程序，再经过一次编译，转换成计算机上可执行的程序。模 拟语言能减轻用户的程序工作，但也不可避免地带来一些限制，消耗更多的内存和运算时间。

模拟语言可分为离散事件模拟语言(如 GPSS 及其各种改型、SIMSCRIPT 、GASD 、 CSL 、SIMULA等)和连续系统模拟语言(如 DARE 、ACSL 、CSS 、CSSL等)两大类型。 对于各种应用领域还有专用的模拟语言。

程序模拟与计算机硬件和软件技术的发展关系非常密切。为便于建立模型和进行模型 的有效性检验，使模拟模型在时间和空间上与真正的系统有一定程度的相似关系。在模拟 过程中，希望能方便地改变参数甚至改变模型的结构，并能通过键盘命令随时输出数据和 图表。因此，程序模拟要求计算机有很强的并行处理能力，有较高的运算速度，有人机交 互能力和便于使用的模拟语言。

程序模拟的尺度既可以是宏观的，也可以是微观的。在宏观尺度上，可以用实验数据 库预测工艺流程、操作条件及系统的性质，计算材料的力学和加工性能， 一般用于化工过 程模拟、机械制造等领域。在微观尺度上，微观粒子的结构和性质起重要作用， 一般用于 反应机理研究、宏观性质模拟等。

**2.4** **科学研究第四范式**

在20世纪，蕴藏着科学理论的科学数据经常被掩埋在零零散散的实验记录中，只有少 数的大项目数据存储在磁介质中。来自单个的、小型的实验室科学数据很容易丢失掉，大 数据管理与支持科研群体获取分布保存的数据成为巨大的挑战。

图灵奖获得者、美国计算机科学家詹姆斯·尼古拉斯·吉姆·格雷于2007年1月11 日在计算机科学与电信委员会上的最后一次演讲中描绘了关于科学研究第四范式的愿景。 这个范式成为由实验、理论与仿真所主宰的早期历史阶段的符合逻辑的自然延伸。

詹姆斯·尼古拉斯·吉姆·格雷博士生于1944年，如图2-4所示。2007年1月28日 在海上失踪。他作为一个程序员、数据库专家、工程师与研究者而广受尊重。1969年吉姆



成为吉利福尼亚大学伯克利分校第一个获得计算机科学博士学 位的人，曾在多个著名高科技公司工作，包括贝尔实验室、IBM 研究院、天腾计算机公司、数字设备公司，最后加盟硅谷的微 软研究院。

如果采用传统的第一、第二、第三范式的研究方法来直接 研究密集型数据本身已经无法进行模拟推演，无法通过主流软 件工具在合理的时间内抽取、处理、管理并整合成为具有积极 价值的服务信息。正是在这样的环境下，提出了科学研究第四 范式，本范式是以数据考察为基础，联合理论、实验和模拟一 体的数据密集计算的范式，数据被捕获或者由模拟器生成，利



第2章 科学研究范式



图2-4 詹姆斯·尼古拉斯 ·

吉姆·格雷博士

用软件处理，信息和知识存储在计算机中，科学家使用数据管理和统计学方法分析数据。

**2.4.1** **数据密集型计算**

数据量的急剧增长以及对在线处理数据能力要求的不断提高，使海量数据的处理问题 日益受到关注。源于自然观测、工业生产、产品信息、商业销售、行政管理和客户记录等 的海量数据在信息系统中所扮演的角色正在从“被管理者”向各类应用的核心转变，并已 经成为企业和机构的最有价值的资产之一。其典型特点是海量、异构、半结构化或非结构 化。通过网络提供基于海量数据的各类互联网服务或信息服务，是信息社会发展的趋势。 这一趋势为业界和学术界提出了新的技术和研究问题。这类新型服务的重要特征之一是它 们都是基于海量数据处理的。在这种背景下，数据密集型计算作为新型服务的支撑技术引 起广泛关注。

**1.** **数据密集型计算的特点**

数据密集型计算是指能推动前沿技术发展的对海量和高速变化的数据的获取、管理、 分析和理解。

数据密集型计算具有下述特点。

① 其处理的对象是数据，是围绕数据展开的计算。需要处理的数据量非常巨大，且变 化快，是分布的、异构的。因此，传统的数据库管理系统不能满足其需求。

② 计算的含义是从数据获取到管理再到分析、理解的整个过程。因此，数据密集型计 算既不同于数据检索和数据库查询，也不同于传统的科学计算和高性能计算，是高性能计 算与数据分析和挖掘的结合。

③ 其目的是推动技术前沿发展，目标是依赖传统的单一数据源和准静态数据库所无法 实现的应用。

**2.** **数据密集型计算的典型应用**

(1)万维网应用

无论是传统的搜索引擎还是新兴的Web 2.0 应用，都是以海量数据为基础，以数据处 理为核心的互联网服务系统。为支持这些应用，系统需要存储、索引、备份海量异构的万



**大数据基础与应用**

维网 (Web) 页面、用户访问日志以及用户信息，并且还要保证能快速准确地访问这些数 据。这需要数据密集型计算系统的支持，因此Web 应用成为数据密集型计算的发源地。

(2)软件即服务应用

软件即服务 (Software as a Service,SaaS) 通过提供公开的软件服务接口，使用户能够 在公共的平台上得到定制的软件功能，为用户节省了软硬件平台的购买和维护费用，也为 应用和服务整合提供了可能。由于用户的各类应用所涉及的数据具有海量、异构和动态等 特性，因此有效地管理和整合这些数据，并在保证数据安全和隐私的前提下提供数据融合 和互操作功能，需要数据密集型计算系统的支持。

(3)大型企业的商务智能应用

大型企业地理上往往是跨区域分布的，互联网为其提供了统一管理和全局决策的平台。 实现企业商务智能需要整合生产、销售、供应、服务、人事和财务等一系列子系统。数据 是整合的对象之一，更是实现商务智能的基础。由于这些子系统中的数据包括产品设计、 生产过程、计划、客户、订单以及售前售后服务等，类型多样，数量巨大，结构复杂和异 构，因此数据密集型计算系统是实现跨区域企业商务智能的支撑技术。

**3.** **数据管理**

数据密集型计算系统中的数据管理问题是核心问题。其与传统的数据管理问题相比， 在应用环境、数据规模和应用需求等方面有本质区别。

数据密集型计算处理的是海量、快速变化、分布和异构的数据，数据量一般是 TB 甚 至是 PB 级的，因此传统的数据存储和索引技术不再适用。地理上的分散性、模型和表示 方式的异构性给数据的获取和集成带来了困难。数据的快速变化特性要求处理必须及时， 而传统的针对静态数据库或者数据快照的数据管理技术已无能为力。

数据密集型计算中“计算”的含义是多元的。它既包括搜索、查询等传统的数据处理， 也包括分析和理解等“智能”处理。数据密集型计算所需要的数据分析和理解不仅仅是单 一的数据分析或挖掘算法，这些算法必须能够在海量、分布和异构数据管理平台上高效地 实现。数据特性决定了不可能为每一个数据分析和理解任务从存储和索引开始开发新的算 法。因此，数据密集计算需要的是与存储和管理平台紧密结合的、具有高度灵活性和定制 能力的、易用的数据搜索、查询和分析工具。使用这一工具，用户可以构造复杂的数据分 析甚至理解应用。由于数据密集型计算要求在海量存储和高性能计算平台上实现，因此数 据密集型计算通常无法在本地提供服务。有效方式是以Web 服务方式提供应用接口。然而 和传统的高性能计算不同，用户的要求可能包括从数据获取到预处理再到数据的分析、处 理的整个过程，可能涉及复杂的流程。因此，数据密集型计算应用的服务接口必须提供整 体流程的描述功能，并提供良好的客户机与服务器之间的基于Web 服务的交互功能。

**2.4.2** **格雷法则**

数据密集型科学由三个基本活动组成，即数据的采集、管理与分析。数据从各种不同 规模和性质的来源涌来，主要有：大型国际实验，跨实验室、单一实验室或个人观察实验，





第2章 科学研究范式

以及个人生活之中。各种实验涉及多学科、大规模数据，特别是它们的高数据通量，对合 适的数据采集、管理与分析工具形成巨大的挑战。

对于大型科学数据集的大数据工程，吉姆·格雷制定了非正式法则或规则，如下所述。

**1.** **科学计算数据爆炸式增长**

科学数据的爆炸式增长为前沿科学的研究带来了巨大挑战，数据的增长已经超过数十 亿字节。因此，对大数据的采集、管理与分析是新的挑战。计算平台的 I/O 性能限制了观 测数据集的分析与高性能的数值模拟，当数据集超出系统随机存储器的能力，多层高速缓 存的本地化将不再发挥作用，仅有很少的高端平台能提供足够快的I/O 子系统。

高性能、可扩展的数值计算也对算法提出了挑战，传统的数值分析包只能在适合RAM 的数据集上运行。为了进行大数据的分析，需要对问题进行分解。通过解决小问题获得大 问题解决的还原论方法，是一种重要方法。

**2.** **解决方案为横向扩展的体系结构**

对网络存储系统进行扩容并将它们连接到计算节点群中并不能解决问题，因为网络的 增长速度不足以应对必要存储逐年倍增的速度。横向扩展的解决方案提倡采用简单的结构 单元，在这些结构单元中，数据被本地连接的存储节点所分割，这些较小的结构单元使得 CPU、磁盘和网络之间的平衡性增强。格雷提出了网络砖块的概念，使得每一个磁盘都有 自己的 CPU 和网络。尽管这类系统的节点数将远大于传统的纵向扩展体系结构中的节点 数，但每一个节点的简易性、低成本和总体性能足以补偿额外的复杂性。

**3. 将计算用于数据而不是数据用于计算**

大多数数据分析以分级步骤进行。首先对数据子集进行抽取，通过过滤某些属性或抽 取数据列的垂直子集完成，然后以某种方式转换成聚合数据。

近年来， MapReduce 已经成为分布式数据分析和计算的普遍范式，其原理类似于分布 式分组和聚合的能力。根据这一原理构造的 Hadoop 开源软件已成为目前大数据处理的最 好的工具，Hadoop 技术成为推动大数据安全计划的引擎。企业使用 Hadoop 技术收集、共 享和分析来自网络的大量结构化、半结构化和非结构化数据。

Hadoop 是一个开源框架，它实现了 MapReduce 算法，用以查询在互联网上的分布数 据。在MapReduce 算法中，映射 (Map) 的功能是将查询操作和数据集分解成组件， Reduce 的功能是在查询中映射的组件可以被同时处理(即约简),从而可以快速地返回结果。

Hadoop 具有方便、健壮、可扩展、简单等一系列特性。Hadoop 处理数据是以数据为 中心，而不是传统的以程序为中心。在处理数据密集型任务时，由于数据规模太大，数据 搬移变得十分困难，Hadoop 强调把程序向数据迁移。也就是说，以计算为中心转变为以数 据为中心。

**4.** **以20个询问规则开始设计**

(1)20个询问规则

20个询问规则是设计步骤的别称，这一步骤使专门领域科学家与数据库设计者可以对





**大数据基础与应用**

话，填补科学领域使用的动词与名词之间，以及数据库中存储的实体与关系之间的语义鸿 沟。这些询问定义了专门领域科学家期望对数据库提出的有关实体与关系方面的精确问题 集。这种重复实践的结果是：专门领域科学家和数据库之间可以使用共同语言。

在“20个询问”开始设计启发式规则中，是指在完成科研项目时，研究人员要求数据 系统回答的20个最重要问题，5个问题不足以识别广泛的模式，100个问题导致重点不突 出。由于与人类选择有关的大多数决定都遵循长尾理论，询问中的相关信息根据重要性排 序显然是呈对数分布。

这种方法非常成功地使设计过程聚焦于系统必须支持的最重要特征，同时帮助专门领 域科学家理解数据库系统中的折中，从而限制系统的蠕动。

(2)长尾理论

长尾理论是网络时代兴起的一种新理论。长尾实际上是统计学中幂律和帕累托分布特 征的一个通俗的表达。过去人们只能关注重要的人或重要的事，如果用正态分布曲线来描 绘，人们只能关注曲线的头部，而将处于曲线尾部或者需要更多的精力和成本才能关注到 的大多数人或事忽略。例如，在销售产品时，厂商关注的是少数几个VIP 客户，无暇顾及 在人数上居于大多数的普通消费者。而在网络时代，由于关注的成本大大降低，有可能以 很低的成本关注正态分布曲线的尾部，但是关注尾部产生的总体效益甚至会超过头部。例 如，某著名网站是世界上最大的网络广告商，它没有一个大客户，收入完全来自被其他广 告商忽略的中小企业。安德森认为，网络时代是关注长尾、发挥长尾效益的时代。

再举例来说， 一家大型书店通常可摆放10万本书，但亚马逊网络书店的图书销售额中， 有四分之一来自排名10万以后的书籍。这些冷门书籍的销售比例正以高速成长，预估未来 可占整个书市的一半。这表明，消费者在面对无限的选择时，真正需要的东西和想要取得 的渠道都出现了重大的变化， 一套崭新的商业模式也跟着崛起。简而言之，长尾所涉及的 冷门产品涵盖了几乎更多人的需求，当有了需求后，将有更多的人意识到这种需求，从而 使冷门不再冷门。

长尾理论是对传统的二八定律的颠覆。人类一直在用二八定律来界定主流，计算投入 和产出的效率。它贯穿了整个生活和商业社会。这是1897年意大利经济学家帕累托归纳出 的一个统计结论，即20%的人口享有80%的财富。当然，这并不是一个准确的比例数字， 但表现了一种不平衡关系，即少数主流的人(或事物)可以造成主要的、重大的影响。以 至于在市场营销中，为了提高效率，厂商们习惯于把精力放在那些有80%客户常购买的20% 的主流商品上，着力维护这些购买其20%商品的80%的主流客户。在上述理论中被忽略不 计的80%就是长尾。

传统的市场曲线是符合80/20铁律的，为了抢夺那带来80%利润的畅销品市场，争夺 激烈。但是互联网的出现改变了这种局面，所谓的热门商品正越来越名不副实，99%的商 品都有机会进行销售，市场曲线中那条长长的尾部(所谓的利基产品)成为可以寄予厚望 的新的利润增长点。

**5.** **工作至工作**

工作至工作是指从工作版本向工作版本的升级，这是一个设计法则。无论数据驱动的



第2章 科学研究范式

计算体系结构变化多么迅速，尤其是当涉及分布数据的时候，新的分布计算模式每年都出 现新的变化，使其很难停留在多年的自上而下的设计和实施周期中。当项目完成之时，最 初的假设已经变得过时，如果要建立只有每个组件都发挥作用的情况下，才开始运行的系 统，那么将永远无法完成这个系统。在这样的背景下，唯一的方法就是构建模块化系统。 随着潜在技术的发展，这些模块化系统的组件可以被代替，现在以服务为导向的体系结构 是模块化系统的优秀范例。

**2.4.3** **核心内容**

科学研究的范式是科学家用于科学研究的范式，而不是科学知识的各种范式。相比库 恩科学动力学理论，网络可以帮助我们更好地理解大数据策略。

**1.** **科学研究范式的演化过程**

在漫长的科学研究范式进化过程中，最初只有实验科学范式，主要描述自然现象，以观 察和实验为依据的研究，又称为经验范式。后来出现了理论范式，是以建模和归纳为基础 的理论学科和分析范式，科学理论是对某种经验现象或事实的科学解说和系统解释，是由 一系列特定的概念、原理(命题)以及对这些概念、原理(命题)的严密论证组成的知识 体系。开普勒定律、牛顿运动定律、麦克斯韦方程式等正是利用了模型和归纳而诞生的。 但是，对于许多问题，用这些理论模型分析解决过于复杂，只好走上了计算模拟的道路，

提出了第三范式。第三范式是以模拟复杂现象为基础的计算科学范式，又可称为模拟范式。 模拟方法已经引领我们走过了20世纪后半期的全部时间。现在，数据爆炸又将理论、实验 和计算仿真统一起来，出现了新的密集型数据的生态环境。模拟方法正在生成大量数据， 同时实验科学也出现了巨大数据增长。研究者已经不用望远镜来观看，取而代之的是通过 把数据传递到数据中心的大规模复杂仪器上来观看，开始研究计算机上存储的信息。

无须质疑，科学的世界发生了变化，新的研究模式是通过仪器收集数据或通过模拟方法 产生数据，然后利用计算机软件进行处理，再将形成的信息和知识存储于计算机中。科学家 通过数据管理和统计方法分析数据和文档，只是在这个工作流程中靠后的步骤才开始审视数 据。可以看出，这种密集型科学研究范式与前三种范式截然不同，所以将数据密集型范式从 其他研究范式中区分出来，作为一个新的、科学探索的第四种范式，其意义与价值重大。

**2.** **数据密集型科学的基本活动**

数据密集型科学由数据的采集、管理和分析三个基本活动组成。数据的来源构成了密 集型科学数据的生态环境，主要有大型国际实验，跨实验室、单一实验室或个人观察实验， 个人生活等。各种实验涉及多学科的大规模数据，例如澳大利亚的平方公里阵列射电望远 镜、欧洲粒子中心的大型强子对撞机、天文学领域的泛 STARRS 天体望远镜阵列等每天能 产生几个千万亿字节 (PB) 的数据。特别是它们的高数据通量，对常规的数据采集、管理 与分析工具形成巨大的挑战。为此，需要创建一系列通用工具支持从数据采集、验证到管 理、分期和长期保存等整个流程。





**大数据基础与应用**

**3.** **学科的发展**

关于学科的发展，格雷认为所有学科都有两个进化分支， 一个分支是计算，另一个分 支是信息。如生态学可以分为计算生态学和生态信息学，前者与模拟生态的研究有关，后 者与收集和分析生态信息有关。在信息学中，把由实验和设备产生的、档案产生的、文献 中产生的、模拟产生的事实以编码和表达知识的方式存在一个空间中，用户通过计算机向 这个空间提出问题，并由系统给出答案。为了完成这一过程，需要解决的一般问题有：数 据获取、管理 PB 级大容量的数据、公共模式、数据组织、数据重组、数据分享、查找和 可视化工具、建立和实施模型、数据和文献集成、记录实验、数据管理和长期保存。可以 看出，科学家需要更好的工具来实现大数据的捕获、分类管理、分析和可视化。

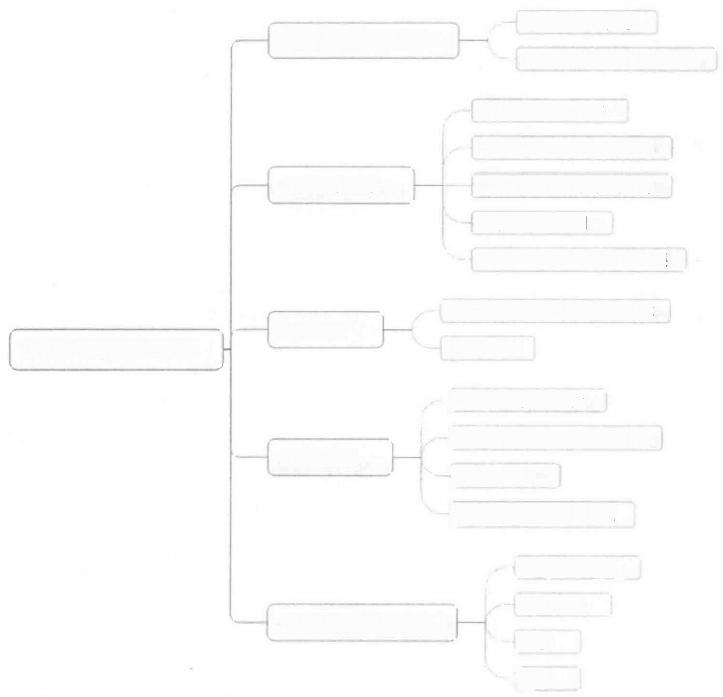
小结

本章介绍了实验科学、理论推演、计算机仿真三种范式，以及科学研究第四范式。 进一步探讨了这种新范式的内涵和内容，包括利用多样化工具不间断地采集科研数据、建 立系统化工具和设施来管理整个数据生命周期，开发基于科学研究问题的数据分析及可视 化工具与方法等。



**第3章** **分布系统设计的CAP** **理论**

本章主要内容



可伸缩性的概念

分布式系统的伸缩性

影响横向扩展的主要因素

可伸缩共享数据库

对等复制的横向扩展方案

链接服务器和分布式查询

分布式分区视图

数据依赖型路由的横向扩展

分布系统设计的核心系统需求

CAP理论

CAP定理

三个核心需求分析

ACID 、BASE与CAP的关系

BASE 模型

CAP 与延迟

CAP理论的进一步研究

系统核心需求

系统服务

Web分布式系统设计

冗余

分区

分布系统设计的CAP理论

横向扩展方案

在计算机科学中， CAP 理论又称为布鲁尔定理 (Brewer's theorem),是由柏克莱加州 大学计算机科学家埃里·布鲁尔 (Eric Brewer) 在1998年提出一个假说，并在2000年的 分布式计算原则研讨会上发表。这个假说是布鲁尔及同事在横向可伸缩性分布系统设计方 面的多年辛勤劳动的结晶。2002年，麻省理工学院的赛斯·吉尔伯特 (Seth Gilbert) 和南 希·林奇 (Nancy Lynch) 又完成了布鲁尔假说的证明，使之脱离了唯象学说而成为一个定 理。但应说明的是，吉尔伯特和林奇证明的布鲁尔定理比布鲁尔提出的假说更为狭义。目 前 ，CAP 理论已成为分布式系统设计与构建的重要理论基石。



大数据基础与应用

**3.1** **分布式系统的伸缩性**

分布式系统的伸缩性是一个重要概念，伸缩性是CAP 理论的基础概念。

**3.1.1** **可伸缩性的概念**

可伸缩性 (Scalability) 又称可扩展性，是指通过扩展系统规模来提高性能并处理更大 数量的用户和通过扩展系统规模提高系统的容错能力等。

在分布系统设计中，利用横向扩展与纵向扩展来实现可伸缩性。

**1.** **横向扩展与纵向扩展**

横向扩展(Horizontal Scaling)是指向逻辑单元之外的扩展，增加更多逻辑单元的资源， 并使它们像一个单元一样工作。大多数集群方案、分布式文件系统、负载平衡等都可以提 高横向的可伸缩性。

纵向扩展(Vertical Scaling) 是指在同一个逻辑单元内增加资源来提高处理能力。例如， 在现有服务器上增加 CPU 计算能力，或者在现有的 RAID/SAN 存储中增加硬盘来提高存 储能力。

更具体地说，横向扩展是在原有的分布系统添加一个新机器。纵向扩展在原有的机器 添加 CPU 、内存等软硬件资源。图3-1所示的是纵向扩展的示意图，而图3-2所示的是横 向扩展的示意图。纵向扩展的优点是在单一系统上扩展，不需要进行复杂的配置。但其副 作用是导致成本的非线性飙升，进而使得大部分网络服务应用都望而却步，而改用横向扩 展方式。

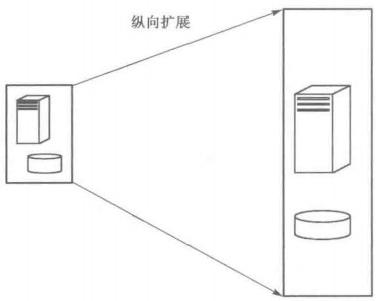
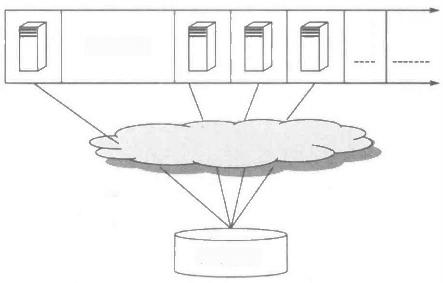


图3-1 纵向扩展示意图



横向扩展

存储网络

共享数据

图3-2 横向扩展示意图

**2.** **其他的扩展性**

除了横向扩展与纵向扩展概念之外，还有下述几种较常用的扩展概念。



第 3 章 分布系统设计的CAP 理论

● 线性扩展性：在扩大规模的时候，扩展因子保持为常数。

● 次线性扩展性：扩展因子小于1.0的扩展。

● 超线性扩展性：因为增加更多组件而获得更佳的性能，在 RAID 系统中跨多个磁 盘的 I/O, 磁盘越多，系统性能越好。

● 负扩展性：当规模扩大的时候，系统性能变坏。

对于应用程序，可扩展性是指应用程序有效使用更多资源以执行更多工作的能力。例 如， 一个在单处理器系统上供给4个用户使用的应用程序可能在一个四处理器系统上服务 于16个用户。在这种情况下，则该应用程序是可扩展的。如果添加更多处理器但却不能增 加所服务的用户数量，则该应用程序是不可扩展的。

可以利用纵向扩展和横向扩展实现系统的可伸缩性。纵向扩展表明扩展到更大、功能 更强的逻辑单元。例如，对于数据库，从4个处理器的服务器扩展到16 或 3 2 处理器的 服务器，这是最常见的数据库扩展方式。如果为了提升数据库系统的功能，除了部署更高 版本的数据库管理系统之外，由于当前硬件资源不够用，还需要增加具有更多处理器和更 多内存的硬件。纵向扩展的优势是：不需要对数据库进行重大更改。通常，只需在较大的 系统中安装数据库管理系统并以常用方式保持运行，就可以使用更多的数据库功能来处理 更多负载。横向扩展表明扩展到多个服务器。横向扩展通常具有可利用初始硬件成本这一 优势，8台4处理器服务器通常比一台32处理器服务器更便宜，但是如果把授权和维护成 本算在内，这项优势往往就被抵消掉了。但是，在某些情况下，从可用性角度看，横向扩 展解决方案还可以提高系统冗余度。

**3.1.2** **影响横向扩展的主要因素**

横向扩展与纵向扩展是一种架构，这些架构可以应用在存储、数据库及网络上等。

应用程序需要处理多种类型数据，但是每种类型数据对横向扩展体系结构要求不同。

如果只有一个数据库，很容易以相同方式处理所有数据，但如果要拆分和复制数据构成分 布数据库，就需要进行横向扩展，就需要了解如何使用各种类型数据，以便选择正确的解 决方案。本节介绍引用数据、活动数据和资源数据3种数据，以及它们对横向扩展的影响。

**1.** **引用数据**

某些数据更新非常频繁。例如，Web 服务器的日志数据以及某个车间的计算机的仪表 读数等。但是，季度销售额这样的历史数据却更新不频繁，甚至不进行更新。引用数据是 仅提供给应用程序使用、不用对其进行维护的数据。引用数据的主要特点是不但相对稳定， 而且在设定时间内有效。例如，订单输入系统使用的产品目录、航班时刻表以及金融系统 中使用的账目表，这种数据相对稳定，由于它可以提供给其他应用程序使用，因此频繁的 更改会导致混乱。如果一个价格列表中的价格在一天内更改多次，则客户会感觉很混乱而 且不满意。引用数据通常以固定的时间间隔进行更改。引用数据还有一个版本标签，该标 签包含在引用该数据的事务中。例如，采购订单可能会被用于创建订单的目录版本，以便 在定价时不引起混乱。业务可能选择接受若干个引用数据的版本，以避免客户使用过期的





**大数据基础与应用**

引用数据。

由于引用数据在一定时期内是稳定的，而且通常有一个版本标识符来标识使用了引用 数据的哪个副本，因此可以将其复制到很多不同的系统，进而避免产生不一致情况。在网 络域中，每个Web 服务器都可能拥有目录的一个副本，以便快速响应目录浏览请求。在很 多情况下，常用的引用数据缓存在内存中。稳定的引用数据可传输到智能客户端，以提供 快速访问和脱机浏览功能。如果该副本丢失或损坏，可以方便地获取一个新副本。引用数 据版本标识符可用于确定当前版本是否可用。所有引用数据副本都是主控副本，因此除非 更改主控，否则这些副本都不更新。这表明不会出现更新冲突问题，因此快照或事务复制 技术可用于更新引用数据。主控副本的所有者可以公开一个服务，以便将引用数据的副本 返回给没有在数据库中存储它的应用程序。上述分析表明，引用数据的横向扩展易于实现， 而且在较少的投资下可提供改善的性能。其他具有引用数据类型相同特点的数据，也可以 使用处理引用数据的方式来处理。例如，历史记录虽然是不断重复的，但很少更改，可以 作为引用数据来处理。

**2.** **活动数据**

活动数据是与特定活动或业务事务相关联的数据。例如，采购订单或股票下跌将生成 一些与该事务相关联的数据。该数据仅在特定业务活动范围内相关，除了一些历史原因外， 当该活动完成之后，这些数据并不是非常有用，活动数据就转变为引用数据。

活动数据的更新率也很低。例如，在创建一个采购订单后，仅当发生状态更改或发货 日期更改时，它才进行更改，但这些事件的发生相对不太频繁，每天仅有几个更新。 一般 而言，活动数据易于识别，并且通常不在活动范围之外访问。这表明，如果活动数据因横 向扩展而在若干个数据库间拆分，将很容易查找。需要访问特定采购订单的业务事务通常 需要知道编号，因此，如果采购订单以编号范围分区，则很容易找到数据库以访问所需的 采购订单。活动数据通常由另一个数据对象限定范围。例如，将特定客户的所有订单存储 在与该客户的客户信息所在的同一数据库中，这是最常见的订单存储方式。同样地，也可 将特定供应商的所有采购订单存储在该供应商所在的供应商数据库中。这表明，在必要时， 复制活动数据相对比较容易，而且在很多情况下，在需要横向扩展时，可以将活动数据分 散到若干个数据库中。采用哪个方法来横向扩展活动数据，主要取决于数据使用率因素。

**3.** **资源数据**

资源数据是与业务相关的核心数据，如库存清单、账户数据、客户档案等都属于资源 数据。如果资源数据丢失，基本上无法进行工作。因此，资源数据库通常使用大量的数据 集成，以高可用性功能来确保资源数据始终可用。资源数据通常具有非常高的并发要求， 因为很多应用程序和用户都需要访问同一数据，需要较高的更新率。库存清单项的可用数 量或账户余额在一天内可能发生多次更改。在满足数据集成和高可用性需要方面，纵向扩 展对资源数据则更适用。

资源数据通常只是当前活动数据。非活动账户、废止的部分等通常在一个相对静态的 历史表中维护，变成了引用数据。资源数据的快照可能由于历史或报告原因才会用到，因





**第3章** **分布系统设计的CAP 理论**

此它们也是引用数据。当资源数据变成引用数据后，对数据集成和高可用性的要求就不重 要。从资源数据到引用数据的转换保留了资源数据的相关性，同时也减少了资源数据的大 小和增加了横向扩展的需要。

**4.** **数据分区**

横向扩展数据最有效的方法之一是将数据分散到多个数据库中，以便每个数据库服务 器可以处理该数据的一部分。虽然这是一个横向扩展数据很直接的方式，但并不是所有数 据都可以有效地进行分区，而且即使它可以分区，其分区方式也将对性能产生较大的影响。 为了说明分区的影响，下面来考虑一个订单数据库进行分区的几种方案。

一是根据订购的内容对订单进行分区，图书订单存储在一个数据库中，服装订单存储 在另一个数据库中等。

二是通过订单号范围来拆分订单，即通过订单号来访问订单。但如果单号与客户表有 大量链接，则该方案也需要分布式链接。解决该链接问题的方法是，按客户号对订单数据 库进行分区。如果客户数据库进行了分区，这将特别有效，而且每个客户的订单与客户位 于同一个数据库中。其他数据可能必须链接到订单数据，如果可能，该数据应该对相同的 方案进行分区以避免分布式链接。该数据的一部分可以是引用数据(如项描述),而且可以 复制到所有订单数据库中，以消除到清单数据库的分布式链接。如果应用程序数据可以划 分到多个数据库中，并且多个服务器提供的额外处理功能优于汇集结果的通信成本，则可 以对该数据进行分区，并不是所有的应用程序数据都可以有效地进行分区，而选择正确的 分区方案对于分区数据的有效横向扩展而言是必要的。

**5.** **数据的相互依赖与耦合**

(1)分布数据的依赖

如果数据库的各个部分由不同的应用程序使用，则以应用程序边界对数据库进行拆分， 这样每个应用程序都可处理自己专用的数据。

如果不同应用程序使用的数据可以进行分段以便提供专用的数据库处理，则不会导致 由于数据相互依赖而引起的过多网络通信量所带来的影响。

拆分数据库的最佳方式就是仔细分析数据模型。“实体—关系”关系图中的关系表示连 接路径和引用完整性约束。跨分布式数据库进行连接和约束强制的开销很高。数据库不会 在数据库之间强制引用约束，因此，如果相关的表在这两个数据库之间进行拆分，就必须 忽略约束，或者必须由触发器或应用程序强制它。例如，在订单项数据库中，通常会发现 客户表和清单表之间存在几项关系。在确定了哪几组表可以在不同的数据库上分隔之后， 则需要查看更新模式。如果在数据库之间拆分数据将导致大量的跨数据库的分布式事务， 则两阶段提交的额外更新开销将抵消一些横向扩展优势。

(2)分布数据的耦合

对于数据耦合需要考虑的最后一个因素是共享表的处理方式。将有一定数量的表由多 个应用程序访问。因此，当拆分数据时，必须确定在何处放置共享表。在某些情况下， 一 个表可能由多个应用程序读取但只由一个应用程序更新，因此使用更新应用程序进行定位





**大数据基础与应用**

时需要合理选择。如果该表相对较小而且广泛用于多个应用程序，则将其复制到多个数据 库很有意义。如果该表由单个应用程序更新，这是最容易的，这样来自主控副本的事务复 制可用于使其他副本保持最新状态。

(3)高耦合资源数据

资源数据由于具有大量完整性约束，所以是高度相互依赖的，这意味着高度耦合。在 某些情况下，不能按应用程序拆分资源数据。解决这个问题的方法是：

● 对应用程序进行某些更改，可以横向扩展引用数据和活动数据；

● 对资源数据进行分区，以便在资源数据高度耦合时仍然可以横向扩展，但是要求 拆分数据的一些横向扩展选项可能要求应用程序重构，以适应数据重构。

**6.** **更改应用程序的能力**

设计应用程序时，需要考虑横向扩展策略，也就是说，横向扩展策略不需要更改应用 程序。早期的应用程序进行横向扩展时，可能需要对查询和存储过程进行少量的更改，或 者可能需要重新考虑应用程序的工作方式。显然，使用的应用程序对横向扩展应具有最大 的灵活性。因此，当设计一个新应用程序时，应该考虑横向扩展。因为如果当应用程序在 生产时耗尽资源后再进行更改，比一开始就设计横向扩展要难得多。任何横向扩展策略对 于应用程序代码必须透明。

从上述的数据类型对横向扩展影响的介绍中可以看出，数据的特征影响对数据的横向 扩展技术的选择。其中，更新现存的数据库管理系统(如 SQL Server 等)都支持若干种横 向扩展技术，而具体使用哪种技术，则取决于所涉及的数据和应用程序的特征。

**3.2** **横向扩展方案**

横向扩展方案有多种，下面介绍几种常用的横向扩展方案。

**3.2.1** **可伸缩共享数据库**

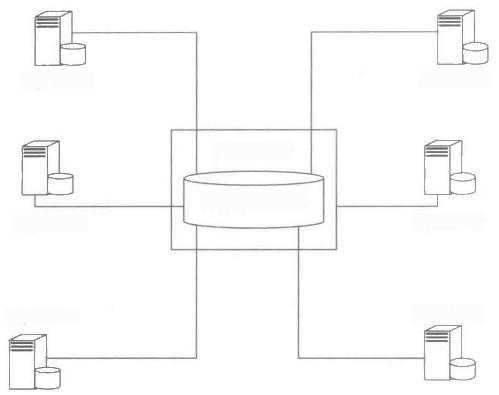
**1.** **更新频率为零的可伸缩共享数据库**

在 SQLServer 中，可伸缩共享数据库是最易于实现的横向扩展方案。在存储区域网络 (SAN) 上创建一个数据库，最多可将8 个运行在不同服务器上的 SQL Server 实例附加 到该数据库，然后开始处理请求。这是传统的共享磁盘的横向扩展方案，在该方案中处理 功能横向扩展，但只使用该数据的单个磁盘图像。每个实例都在其内存中保留它自己的数 据库锁，所有实例都不知道其他实例的锁。这表明可伸缩共享数据库非常适用于数据仓库 或报告数据库，但并不适用于更新数据的应用程序。对于所常用的数据特征而言，仅当更 新频率为零时，可伸缩共享数据库才正常工作。根据定义，该数据是历史数据，因此都是 引用数据。图3-3说明将可伸缩共享数据库用作横向扩展解决方案。





第3章 分布系统设计的 CAP 理论



SQL Server SQL Server

存储域网络

只读数据库文件

SQL Server SQL Server

SQL Server SQL Server

图3-3 可伸缩共享数据库

显然，从未进行更改的数据库其价值有限，因此要更新数据库，所有的 SQL Server 实 例都应该分离该数据库，将其中一个实例以读写模式附加，并使用当前数据刷新该数据库。 如果大量数据已经更改，这一操作可能要花一段时间。因此，如果 SAN 有足够的可用空 间，维护两个数据库是有意义的：在一个数据库更新时，可以使用另一个数据库。

**2.** **更新频率很低的可伸缩共享数据库**

仅当更新频率很低时，可伸缩共享数据库才实用，因为数据库在共享时根本无法更新。 可伸缩共享数据库不需要应用程序更新(假设该应用程序不试图更新该数据库)。该数据库 在不进行更新的情况下横向扩展，因此分区和耦合不是决定使用可伸缩共享数据库的因素。 总之，可伸缩共享数据库在数据仓库、数据中心和报告数据库的应用程序中是有用的，其 中数据更新频率可以减少至周期性批更新。如果该更新模式可用于该应用程序，可伸缩共 享数据库是首选的横向扩展方法，因为它们易于实现且需要最少的应用程序更改。

**3.2.2** **对等复制的横向扩展方案**

横向扩展更新数据(但更新频率相对较低)是通过复制来完成。复制用于将更改传播 给该数据的所有副本，图3-4说明横向扩展方案中的对等复制方法。

对等复制可以将对数据的任何副本进行更改后传播至所有其他副本。对等复制不提供 冲突解决方案，因此仅在只更新给定数据元素的一个副本的配置中使用。例如，如果将对 等复制应用于维护餐饮连锁店清单，则只允许拥有清单项的连锁店更新。这表明，即使所 有连锁店都能看到其他连锁店的清单，但它们只能更改自己的清单。只允许数据项所有者 更新数据项的规则称为数据管理。管理是避免数据更新冲突非常有用的方式，在不适用数 据管理的情况下，可以使用合并复制来处理冲突。因为合并复制必须处理冲突，所以它比 对等复制需要更多的开销，如果可以避免冲突，则更适于使用对等复制。对等复制需要在



**大数据基础与应用**

数据库的每个副本与其他任何副本之间建立复制，因此当涉及多个数据库时，对其进行管 理就成为一个负担。如果对数据库的更新仅限于数据库的单个副本，则使用事务复制以保 持其他副本最新的单个主控副本是最简单且有效的解决方案。

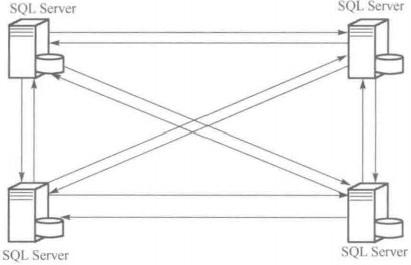
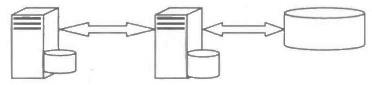


图3-4 对等复制

对于更新频率数据，复制是一个较好的横向扩展方案。数据管理可用于消除更新冲突， 以便可以使用对等复制。如果需要，可以复制数据库中的所有数据，因此当数据不易分区 或者具有高度耦合性时，它非常有用。在大多数情况下，复制不需要更改应用程序，因此 它可用于轻松地横向扩展现有应用程序。复制可在单个表甚至表的某些部分上进行，因此 复制对于横向扩展部分应用程序数据很有用。例如，可以复制目录和价格表这样的引用数 据以改进横向扩展，即使资源数据不横向扩展也是如此。复制在与其他横向扩展解决方案 合并时也很有用。如果通过将活动数据分布到不同数据库来对其进行横向扩展，则可将该 活动数据使用的引用数据复制到所有活动数据库。通常，复制是最简单且应用最广的横向 扩展解决方案之一。

**3.2.3** **链接服务器和分布式查询**

对于应用程序而言，横向扩展的数据库像单个大型数据库一样。对 SQL 查询的主要更 改是，表名将包含数据所在的链接服务器的名称。如果难以更改应用程序，这将使链接服 务器成为一个有吸引力的横向扩展选项，这样用于定位本地表的查询时就可以定位远程表， 而无须更改查询。图3-5说明如何使用链接服务器作为横向扩展解决方案。



本地服务器 **链接服务器**

图3-5 链接服务器

**1.** **具有链接服务器的数据库**

在将数据按功能区域拆分到具有极少耦合的数据库时，最适合使用链接服务器。引用 完整性约束无法使用远程表，因此需要最小化本地数据和远程数据之间的关系。远程查询



**第** **3** **章** **分布系统设计的CAP 理论**

比本地查询的开销大，本地表和远程表之间连接的开销也很大，而且远程表的更新需要分 布式事务。因此，要成功地横向扩展具有链接服务器的数据库，需要仔细设计数据库以最 小化远程数据访问。如果应用程序使用彼此之间具有弱耦合的数据岛，而且跨这些数据 岛的查询也相对较少，则这些数据岛可以在数据库之间分布，从而成为一个有效的横向 扩展解决方案。

**2.** **远程访问数据库**

通过设计横向扩展的数据库，使任何应用程序所用的大部分数据在一个数据库中。例 如，如果某些应用程序在查询中使用客户数据和订单数据，即使该数据是松耦合的，则将 客户数据和订单数据划分到单独的数据库中是无意义的工作。因为无论应用程序打开哪个 数据库，都会继续远程访问数据库的数据。另外，如果其中一些应用程序专门访问客户数 据，其他应用程序访问订单数据而很少访问客户数据，则该划分是有效的。此外，如果几 个查询广泛使用来自这两个数据库的数据，则可以使用复制来减少网络通信量。例如，如 果订单需要有效的客户号，而且每个订单包括一个客户名，则可以将客户数据库的客户号 和客户名列复制到订单数据库。

与复制类似，链接服务器在作为其他横向扩展解决方案的一部分使用时也很有价值。 在任何分布式数据库中，将有一定数量的查询必须跨数据库边界，而链接数据库是支持这 些查询的简单方式。

**3.** **基于类型拆分数据时链接服务器的作用**

基于类型拆分数据时，链接服务器也很有用。例如，如果将60 天以上的历史数据传 输到其他数据库，则链接服务器如同在本地一样允许应用程序访问历史引用数据的查询。 链接服务器解决方案可有效用于某些联机事务处理 (Online Transaction Processing,OLTP) 应用程序。链接服务器不需要或需要少量应用程序更改，因此它们适用于较低版本的应用 程序。确定链接服务器横向扩展解决方案是否适合的关键因素是数据耦合。如果耦合度较 高，则分布式查询的开销将抵消横向扩展带来的性能增益。

**3.2.4** **分布式分区视图**

分布式分区视图能够提供对分区数据的透明横向扩展。例如，根据客户号进行分区： 第一个数据库中是1～10000,第二个数据库中是10001～20000,第三个数据库中是20001~ 30000。SQL Server可以通过查看约束得知哪些客户位于哪个数据库中，这样就可以实现访 问特定客户号的查询仅在包含所需客户号的数据库上执行。不引用特定客户号的查询必须 运行在所有分区上。例如，如果需要查找关于张三的信息，但不知道他的客户号，则 SQL Server 必须查看包含客户表分区的所有数据库。同样地， 一个客户号的两个表之间的连接 可以在每个数据库上并发执行，结果返回至产生查询的数据库。



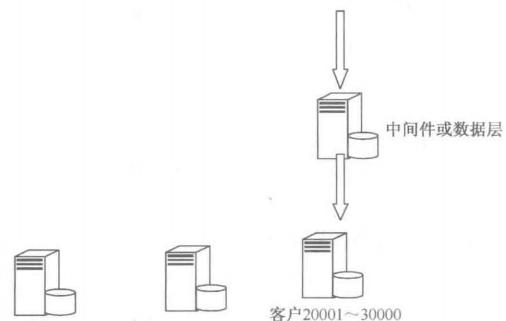


**大数据基础与应用**

**3.2.5** **数据依赖型路由的横向扩展**

通过使用分布式分区视图，SQL Server 可以知道数据分区，并确定在何处找到数据。 使用数据依赖型路由(Data-Dependent Routing,DDR), 应用程序或某些中间件服务通过查 询路由得到正确的数据库。对于应用程序而言， DDR 缺少透明性，但如果该应用程序涉及 关于在何处执行查询决策，则该应用程序需要该数据的更深入的知识，从而促进更好的决 策。如果相同数据有多个可用副本，则 DDR 可以跨副本，进行分布选择，同时将这些更新 指向单个主控副本。将基于请求定位的数据进行路由请求称为数据依赖型路由。数据依赖 型路由的横向扩展方案如图3-6所示。

刷新客户23333



客户10001～20000

客户1～10000

图3-6 数据依赖型路由

多数具有大型事务量的最大的数据库使用 DDR, 在横向扩展中，所有查询将必须指向 单个服务器，因此必须设计数据模型，以便一个查询或更新需要所有数据位于同一服务器 上。例如， 一个较大的联机零售商可能按照客户ID 对数据库进行分区。这意味着关于该客 户的所有信息(包括该客户下的所有订单、客户的账单状态、信用卡信息等)和该客户的 客户数据存储于同一数据库中。

虽然对成千上万的数据库服务器进行大型横向扩展并不常用 DDR, 但是使用相同原则 横向扩展至几十个数据库服务器对于很多应用程序而言是可行的。重定位数据、管理复制 和提取汇总数据等使该解决方案在管理时相对复杂，但该工作的大部分是重复的而且可以 自动进行。

对数据库横向扩展的讨论有一点很重要，那就是任何应用程序都包含不同类型的数据， 而且一个有效的横向扩展解决方案可根据不同数据包含不同的方法：引用数据可在很多不 同位置进行复制和缓存；历史数据可通过对低成本、高容量存储的分布式查询进行公开； 活动数据可跨多个服务器进行分区；资源数据可按应用程序进行拆分。使用横向扩展解决 方案的决策受多个因素的影响。表3-1 比较了这些因素对每个解决方案的重要程度。





第3章 分布系统设计的CAP 理论

**表3-1** **影响横向扩展解决方案选择的因素**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **数据库横向扩展方案** | **更** **新** **频** **率** | **更改应用程序的能力** | **数据可分区性** | **数** **据** **耦** **合** |
| 可伸缩共享数据库 | 只读 | 需要少量更改或  无须更改 | 无要求 | 无要求 |
| 对等复制 | 通常为读取，无 冲突 | 需要少量更改或  无须更改 | 无要求 | 无要求 |
| 链接服务器 | 最小化跨数据库 更新 | 最少更改 | 通常不需要 | 具有低耦合性，非 常重要 |
| 分布式分区视图 | 可以频繁更新 | 可能需要一些更改 | 非常重要 | 影响极小 |
| 数据依赖型路由 | 可以频繁更新 | 可能会进行重大更改 | 非常重要 | 低耦合可能有助于 某些应用程序 |
| 面向服务的数据体系 结构 | 可以频繁更新 | 需要大量更改 | 通常不需要，除非 与 D D R 合 并 | 要求服务之间具 备低耦合性 |

横向扩展体系结构也可以合并多个解决方案。DDR 和 SOA 可以有效地合并，复制和 链接服务器通常是所有横向扩展体系结构的一部分。通过了解应用程序的数据、要求和约 束，可以设计有效的 SQL Server解决方案来满足几乎任何级别的横向扩展。即使横向扩展 不是应用程序的首要要求，但随着应用程序和业务的增长，程序更注重横向扩展设计。

**3.3** **CAP理论**

**3.3.1** **分布系统设计的核心系统需求**

在分布式的环境下，设计和部署横向扩展分布系统时，主要考虑下述三个重要的核心 系统需求。

● 一致性 (Consistency): 所有节点在同一时间具有相同的数据。

● 可用性(Availability): 保证每个请求的成功或者失败都有响应。

● 分区容错性(Partition Tolerance): 系统中任意信息的丢失或失败不影响系统的继 续运行。

通常将上述三个重要的核心系统需求简称为 CAP 需求。

**1.** **一致性**

在分布式系统中数据存在多个副本，对数据的一致性操作必须是原子性的操作，导致 客户对数据的操作(增、删、改)结果是对所有的数据副本全部成功或全部失败，退回一 致性操作的前一状态。 一致性是指数据一致性，其含义为一个服务是一个一致的完整操作。 如果一个存储系统可以保证一致性，那么客户读写的数据完全可以保证是最新的数据，不 会发生两个客户端在不同的存储节点中读取到不同数据副本的情况。例如，在购书的例子



**大数据基础与应用**

中，如果库存中只有一本书，当天只有一个人能买到它。如果两个客户都可以完成订单流 程(如完成支付),那么不一致性就将出现，系统中记录卖出两本书，但仓库中仅存有一本。 从这个例子中可以看出，如果将其扩大到数千个不一致性，并且涉及金融交易中关于买卖 的东西和交易记录的内容不一致，就将发生大问题。对于这个问题，可以利用数据一致性 方法来解决。在购书的订单流程中的某个点减少某本书的库存记录，当其他的客户到达这 个点的时候，书架空了，订单流程将会通知客户，而不进行到支付环节。这样第一个操作 顺利完成，第二个操作则不能完成。数据库非常适合这种情况，因为数据库关注 ACID 属 性，并且通过隔离性来保证一致性，当第一个客户使得库存记录减1,同时使购物车的记 录加1,任何中间状态与第二个客户隔离，第二个客户必须等待几百毫秒，以便数据存储 达到一致状态。 一致性又称为原子性或者事务性。表示一个事务的操作是不可分割的，要 不然这个事务完成，要不然这个事务不完成，不会出现这个事务完成了一半的情况。这种 事务的原子性使得数据具有一致性。

通常情况下，在数据库中存在的脏数据就属于数据没有一致性的表现。而在分布式系 统中，经常出现的一个数据不具有一致性的情况是读写数据时缺乏一致性。比如，两个节 点数据冗余，第一个节点有一个写操作，数据更新以后没有有效地使第二个节点更新数据， 在读取第二个节点的时候就会出现不一致的问题。传统的 ACID 数据库是很少存在一致性 问题的，因为数据的单点原因，数据的存取又具有良好的事务性，不会出现读写的不一致。

最后看一下满足可用性和分区容错性的情况。满足可用性说明数据必须要在不同节点 中有回滚 (replica), 然而还必须保证在产生分区的时候操作仍然可以完成。那么,必然无 法保证操作一致性。

**2.** **可用性**

可用性是指在客户端想要访问数据的时候，可以得到满意的响应，但是，系统可用并 不代表存储系统的所有节点提供的数据完全一致。例如，客户端想要读取文章评论，存储 系统可以返回客户端数据，但是评论缺少最新的一条。这种情况下，仍然说明系统可用。 往往可对不同的应用设定一个最长响应时间，超过这个响应时间的服务可以称之为不可用。 可用性只表明服务是可用的，主要标志是可以完成或不可以完成如上的操作。当购书时希 望得到反馈，而不是浏览器报告网站无法连接的信息。因为网站压力最大，网站通常在业 务最繁忙的时刻挂掉， 一个无法访问的服务没有价值。

好的可用性主要是指系统能够很好地为用户服务，不出现用户操作失败或者访问超时 等用户体验不好的情况。通常情况下，可用性和分布式数据冗余、负载均衡等有着很大的 关联。可用性要求从客户端接收请求的任一节点必须被响应，即使任意的消息可能没有被 正确地发送。如果选择分区容错性和一致性，那么即使节点坏了，操作必须又一致，也能 顺利完成。所以就必须100%保证所有节点之间有很好的连通性。这是很难做到的。最好的 办法就是将所有数据放到同一个节点中，但是显然这种设计是不满足可用性的。

**3.** **分区容错性**

为了提高服务质量，同一个数据的副本存放在不同地点是一种常用的方法。分区容错





**第3章** **分布系统设计的** **CAP 理论**

性是指除了全部网络节点都出现故障之外，所有子集合节点的故障都不能导致整个系统不 正确响应。如果将数据分布在不同的节点上，就有形成分区的风险。假定网线被切断，分 区就形成，节点 A 无法和节点B 通信，无法保证数据一致性。

分区容错性和扩展性紧密相关。在分布式应用中，可能因为一些分布式的原因导致系 统无法正常运转。好的分区容错性要求，应用虽然是一个分布式系统，而看上去却好像是 一个可以运转正常的整体。例如，现在的分布式系统中有某一个或者几个机器坏了，其他 剩下的机器还能够正常运转满足系统需求，这样就具有好的分区容错性。

为了定义分区容错性，假定网络满足如下条件：网络可能丢失从一个节点发往另一个 节点的任意消息，当网络被分区(隔断)时，所有从一个分区的节点发往另一个分区的消 息将会丢失。如果要满足可用性和一致性，那么,为了保证可用，数据必须要有回滚。这 样，系统显然无法容错分区。当同一数据的两个副本分配到了两个无法通信的分区上时， 显然会返回错误的数据。也即，分区容错性是指系统中的数据分布性的大小对系统的正确 性、性能的影响(一定程度上就是可扩展性)。

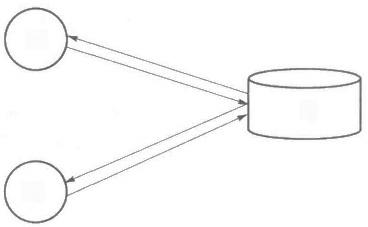
**3.3.2** **CAP定理**

CAP 定理的主要内容是： 一个分布式系统不可能同时满足一致性、可用性和分区容错 性三个系统需求，最多只能同时满足两个系统需求。在考虑满足系统需求时，要根据实际 需要来选择关注点，进而采用相应的策略。CAP 定理明确了实现分布式系统的局限性。

对大型网站，可用性与分区容错性的优先级要高于数据一致性，所以选择了可用性与 分区容错性，然后通过其他方法来保证一致性的商务需求。架构设计师不要将精力浪费在 如何设计能同时满足三者的完美分布式系统，社交SNS 网站可以容忍相对较长时间的不一 致，不影响交易。

**1.CAP 定理的解释**

下面通过一个简单的图示说明 CAP 定理，如图3-7所示。

N₁

V₀

N₂

图3-7 网络中的两个节点

图中的 N₁ 、N₂为网络中的两个节点。它们共享同一数据 V, 其值为V₀。N₁上有一个算 法 A, 可以认为A 中无bug, 并且是可预测和可靠的。N₂ 上有一个类似的算法 B。在这个 例子中，利用A 算法将新值写入 V, 而利用B 算法读取 V 的值。





**大数据基础与应用**

正常情况下的过程如下。

① A 写入新的 V值，称作 V₁。

② N₁发送信息给N₂, 更新V的值。

③ B 读取的 V值将会是 V₁ (见图3-8)。

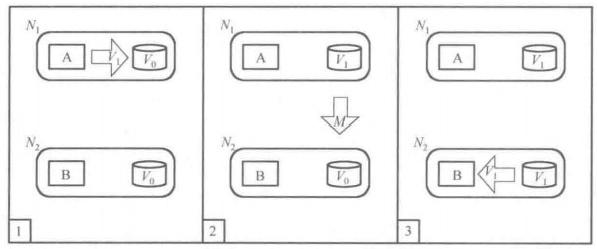


图3-8 正常情况下的过程

如果网络断开(分区),表明从 N₁ 无法发送信息到 N₂, 那么在第(3)步的时候， N₂ 就会具有一个与N₁ 不一致的 V 值，其中在N₁ 中的V=V₁, 在 N₂ 中 的V=V₀。

如果规模达到几百个事务，这将成为一个大问题。如果 M 是一个异步消息，那么 N₁ 无法知道N₂ 是否收到了消息。即使M 是保证能发送的(保证传送),N₁ 也无法知道消息是 否由于分区事件的发生而延迟，或由于N₂ 上的其他故障而延迟。如果M 是一个同步消息， 那将使得N₁ 上 A 的写操作和N₁到 N₂ 的更新事件成为一个原子操作，而这将导致等待问题。 已经说明使用其他的变种方式，即使是部分同步模型也无法保证原子性。因此，CAP 定理 表明，如果使A 和B 高可用(以最小的延迟提供服务),并且使所有的N₁ 到 N₂(n 的值可以 是数百甚至是上千)的节点能够网络分区，那么有时就可能出现某些节点V的值是V₀, 而其 他节点 V 的值是V₁ 。 也就是说，选择了可用性和分区容错性，牺牲了一致性(见图3-9)。

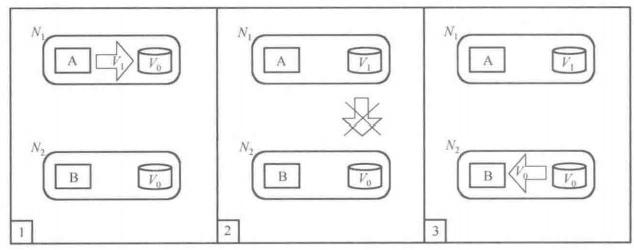


图3-9 分区情况下的过程

**2.CAP 定理的证明**

(1)异步网络模型

在异步网络模型中，没有统一时钟，所有节点仅根据接收到的消息和本地的计算进行 决策。





第3章 分布系统设计的CAP 理论

**定理1** 在一个异步网络模型中，对于所有对等运算(包括消息会素朱的),不可能实 现一个满足以下条件的读写数据对象：

● 可用性。

一致性。

分区容错性。

**证明：** 假设存在一个算法A 能够满足一致性、可用性和分区容错性。构造一次 A 的执 行，包括一个返回非一致结果的请求。假设网络包含至少两个节点，那么它可以 被分为不相关的非空集合：{G,H}。假设所有G 和H 之间的通信消息都丢失(这 是可能的),如果这时在G 上有一个写操作，接着H 上有一个读操作，那么读操 作将无法返回早些的写操作。

推论 在一个异步网络模型中，不可能实现一个满足以下属性的读写数据对象：

● 可用性，所有对等运算。

● 一致性，所有对等运算，但消息不会丢失。

● 分区容错性。

**证明：**主要问题是在异步网络模型中， 一个算法没有办法去判断一个消息是否丢失或者 在传输通道中被延迟。因此，如果在运算中不会丢失任何消息的前提下存在一个 能够保证一致性的算法，那么该算法也能够在所有运算(消息可能丢失)情况下 保证一致性。这与定理1矛盾。

(2)部分同步网络模型

假设一个部分同步的网络模型，所有的节点都有一个时钟，并且所有的时钟以一个相 同的周期计时。但是，这些时钟并不是同步的，在相同的时间，它们显示不同的时间值。 事实上，时钟扮演计时器的角色：处理器可以根据本地状态变量去衡量流逝了多少时间。 一个本地的计时器可以用来调度某事件之后的多长时间间隔进行另一个操作。进一步地， 假设每一个消息要么在给定的时间s 内到达，要么丢失，并且所有的节点在给定时间t 内 处理完一个接收到的消息。

**定理2** 在一个部分同步网络模型中，对于所有对等运算(包括消息会丢失的),不可 能实现一个满足以下属性的读写数据对象：

● 可用性。

● 一致性。

● 分区容错性。

**证明：**证明方法与定理1一样.

但是在部分同步模型中，类似于异步模型推论的结论就不存在了，因此推论的假设基 于节点无法判断一个消息是否丢失。而在部分同步模型中，存在部分同步算法可以在所有 消息传送正常的情况下返回一致性的数据，而仅仅在消息丢失时返回非一致性数据。对于 读或写请求，节点发送一个消息给另一个节点，如果消息返回了，那么节点发送请求的数





大数据基础与应用

据；如果消息在给定的2s+t 时间内没有返回，那么该节点断定消息丢失了，节点就可能返 回一不不一致的请求数据。

**3.CAP 选择**

在实现一个分布式系统时(包括分布式数据库),是不可能同时完美地实现三个方面的。 目前的很多分布式系统是基于首要满足可用性和分区容错性而设计。 一般来说，当处理CAP 的问题时，存在下述三个选择。

(1)放弃分区容错性

为了避免分区问题发生，就必须阻止其发生。 一种做法是将所有的东西(与事务相关 的)都放到一台机器上，但却无法100%地保证分区问题不会发生，因为还是有可能部分失 败，但不太可能碰到由分区问题带来的负面效果。当然，这个选择会严重影响规模限制。

满足一致性、可用性的系统，放弃分区容错性，导致可扩展性不强。放弃分区容错性 的典型系统如下所述：

● Postgres 、MySQL关系数据库；

● Vertica 列/表数据库；

● Aster Data 关系数据库；

● Greenplum 关系数据库。 (2)放弃可用性

相对于放弃分区容错性，其反面就是放弃可用性。 一旦遇到分区事件，受影响的服务 需要等待数据一致，因此在等待期间就无法对外提供服务。在多个节点上控制这一点会相 当复杂，而且恢复的节点需要处理逻辑，以便平滑地返回服务状态。

满足一致性、分区容错性，放弃可用性的系统，通常性能不是特别高。放弃可用性的 典型系统如下所述。

● Bigtable 列/表数据库；

● Hypertable 列/表数据库；

● HBase 列/表数据库；

● MongoDB 文档数据库；

● Terrastore 文档数据库；

● Redis 键值数据库；

● Scalaris 键值数据库；

● MemcacheDB 键值数据库；

● Berkeley DB 键值数据库。 (3)放弃一致性

接受事情会变得最终一致。许多不一致性并不比想到的需要更多的工作(意味着持续 的一致性或许并不是所需要的)。在购书的例子中，如果只有一本库存的书接到了两个订单， 第二个就会成为备份订单。只要告知客户这种情况，请记住这是一种罕见的情况，也许每 个人都会高兴的。

满足可用性、分区容错性的系统，放弃一致性，通常对一致性要求低一些。放弃一致





第3章 分布系统设计的 CAP 理论

性的典型系统如下所述。

● Dynamo 键值数据库；

● Voldemort 键值数据库；

● Tokyo Cabinet 键值数据库；

● KAI 键值数据库；

● Cassandra 列/表数据库；

● CouchDB 列数据库；

● SimpleDB 列数据库；

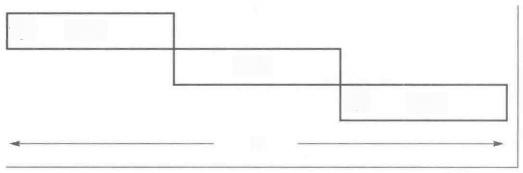
● Riak 列数据库。

**3.4** **BASE** **模型**

最终一致 (Basically Available,Sof-state,Eventually consistent,BASE) 是 ACID 的反 面，但任何架构都不完全基于BASE 或完全基于ACID 。BASE 理论是 ACID 理论与实际相 结合的产物。BASE 在英语中有“碱”的意思，正好和ACID 的“酸”的意义相对。BASE 要求牺牲高一致性，获得可用性或可靠性。

**3.4.1** **三个核心需求分析**

随着互联网应用的飞速发展，数据量与日俱增，传统的 ACID 数据库已经不能满足如 此大的数据存储了。这个时候需要设计出好的分布式数据存储方式。而这些分布式数据存 储方式受到CAP 理论的约束，不可能同时达到高一致性、高可用性、高分区容错性的完美 设计。所以在设计的时候要适当取舍，重点关注对应用需求来说比较重要的，而放弃次要 的，在CAP 的三个核心系统需求之间进行取舍，设计出满足实际应用需求的存储方案。目 前众多的分布式数据系统通过降低一致性来换取可用性。下面是一个简单的例子，可以从 事务的角度分析，如图3-10所示。



@ a;:写

synch

a₂ 读

a



图3-10 基于事务的分析

两个节点数据冗余，第一个节点先有一个写操作，第二个节点后有一个读操作。 a 是 整个过程，要具有一致性，则需要等待a₁ 进行写，然后同步到a₂, 然后a₂ 再进行写，只有 整个事务完成以后，a₂ 才能够进行读。这使得整个系统的可用性下降。 a₂ 一直阻塞在那里 等待a₁ 同步到a₂ 。这个时候如果对一致性要求不高， a₂ 可以不等待 a₁ 数据对于a₂ 的写同





**大数据基础与应用**

步，直接读取，这样虽然此时的读写不具有一致性，但是在后面可以通过异步的方式使得 a₁ 和a₂ 的数据最终一致，达到最终一致性。

在小规模、低压力、小延迟的情况下， CAP 定理还不足以对分布系统总体的性能造成 影响，但随着活动的增加与吞吐量的上升将产生错误。

**3.4.2** **ACID、BASE** **与** **CAP** **的关系**

ACID 和 BASE 代表了两种截然相反的设计思想，ACID 注重一致性，是数据库的传统 设计思路。BASE 在20世纪90年代后期被提出，抓住了当时正逐渐成形的高可用性的设 计思路，并且把不同性质之间的取舍进行研究。大规模跨区域分布的系统，包括云在内， 同时运用了这两种思路。出现较晚的BASE 是基本可用、软状态、最终一致性的英文缩写。 其中的软状态和最终一致性这两种技巧善于分区的场合，并提高了可用性。

CAP 与 ACID 的关系复杂，也更易引起误解。其中一个原因是 ACID 中的C 和 A 所代 表的概念不同于CAP 中的C 和 A。 还有一个原因是选择可用性只部分地影响 ACID 约束。 ACID 四项特性如下所述。

**1.** **原子性** **(A)**

所有的系统都基于原子性操作。当考虑可用性的时候，没有理由去改变分区两侧操作 的原子性，而且满足ACID 定义的、高抽象层次的原子操作，实际上会简化分区恢复。

**2.** **一致性** **(C)**

ACID 的 C 指的是事务不能破坏任何数据库规则，例如，键的唯一性。与之相比，CAP 的 C 仅指单一副本这个意义上的一致性，因此只是ACID 一致性约束的一个严格的子集。 ACID 一致性不可能在分区过程中保持，因此分区恢复时需要重建ACID一致性。推而广之， 分区期间也许不可能维持某些不变性约束，所以有必要仔细考虑哪些操作应该禁止，分区 后又如何恢复这些不变性约束。

**3.** **隔离性** **(**1)

隔离是CAP 理论的核心，如果系统要求 ACID 隔离性，那么它在分区期间最多可以在 分区一侧维持操作。事务的可串行性要求全局的通信，因此在分区的情况下不能成立。在 分区恢复时进行补偿，在分区前后保持一个较弱的正确性定义是可行的。

**4.** **持久性** **(D)**

牺牲持久性没有意义，理由和牺牲原子性一样，虽然开发者有理由(持久性成本太高) 选择 BASE 风格的软状态来避免实现持久性。分区恢复可能因为回退持久性操作，而无 意中破坏某项不变性约束。但只要恢复时给定分区两侧的持久性操作历史记录，破坏不 变性约束的操作还是可以被检测出来并修正的。 一般来说，让分区两侧的事务都满足 ACID 特性会使得后续的分区恢复变得更容易，并且为分区恢复时事务的补偿工作奠定基 本的条件。





第3章 分布系统设计的CAP 理论

**3.4.3** **CAP 与延迟**

虽然CAP 理论的经典解释忽略网络延迟，但在实际中延迟和分区密切相关。CAP 理论 实际应用的场景发生在操作的间歇，系统需要在这段时间内做出关于分区的选择。

● 如取消操作，则降低系统的可用性；

● 如继续操作，则可能损失系统的一致性。

依靠多次尝试通信的方法可以达到一致性。比如，两阶段事务提交，仅仅是推迟了决 策的时间，系统终究要做出一个决定，无限期地尝试下去，本身就是选择一致性而牺牲可 用性的表现。因此以实际效果而言，分区相当于对通信的时限要求。如果系统不能在时限 内达到数据一致性，就表明发生了分区的情况，必须就当前操作在C 和 A 之间做出选择。 从这个实用的观察角度出发可以推导出下述的结论。

● 有些节点检测到了分区，有些可能没有。

● 检测到分区的节点，即进入分区模式，这是优化C 和 A 的核心环节。

可以根据期望中的响应时间，有意识地设置时限。时限设得越短，系统进入分区模式 越频繁，其中有些时候并不一定真的发生了分区的情况，可能只是网络变慢而已。在跨区 域的系统，有时候放弃强一致性来避免高延迟是非常有意义的。Yahoo 的 PNUTS 系统因为 以异步的方式维护远程副本而带来数据一致性的问题。但好处是主副本放在本地，减少操 作的等待时间。这个策略在实际中很实用，因为一般来讲，用户数据大都根据用户的地理 位置做分区。最理想的状况是每一位用户都在他的数据主副本附近。Facebook 使用了相反 的策略：主副本被固定在一个地方，因此远程用户一般访问到的是离他较近，但可能已经 过时的数据副本。不过当用户更新其页面的时候是直接对主副本进行更新，而且该用户的 所有读操作也被短暂转向从主副本读取，尽管这样延迟比较高。在20s 后，该用户的流量 被重新切换回离他较近的副本，此时副本应该已经同步完成了刚才的更新。

**3.4.4** **CAP 理论的进一步研究**

基于BASE 理论建立的分布系统能够满足下述三个系统核心需求。

● 高一致性：读取的数据都是基于快照的，而且错误的更新操作不会执行。

● 高可用性：读取和更新都会返回数据。

● 高分区容错性：允许网络或者节点出错。

如果关注的是一致性，那么就需要处理因为系统不可用而导致的写操作失败的情况。 而如果关注的是可用性，那么应该知道系统的读操作可能不能精确地读取到写操作写入的 最新值。因为系统的关注点不同，相应地采用的策略也是不一样的，只有真正地理解了系 统的需求，才有可能利用好CAP 理论。而对大型网站，可用性与分区容错性优先级要高于 数据一致性， 一般会向注重 A 、P 的方向设计，然后通过其他手段保证对于一致性的商务 需求。架构设计师不应把精力浪费在能满足三者需求的完美分布式系统设计上，而是应该 进行取舍。





大数据基础与应用

为了更好地描述与分析客户端的一致性，通过以下的场景来进行说明，该场景中包括 三个组成部分。

● Process A: 主要实现对存储系统的读写操作。

● Process B和 Process C:Process B和 Process C 与 Process A 独立，并且 Process B 和 Process C 也相互独立，它们同时也可实现对存储系统的读写操作。

下面利用以上场景来描述不同程度的一致性。

(1)强一致性

假如 Process A先写入了一个值到存储系统，存储系统保证后续A 、B 、C 的读取操作 都将返回最新值。

(2)弱一致性

假如 Process A先写入了一个值到存储系统，存储系统不能保证后续A 、B 、C 的读取 操作能读取到最新值。此种情况下有一个“不一致性窗口”的概念，它特指从 A 写入值， 到后续操作 A、B、C 读取到最新值这一段时间，出现了延迟处理，即 A 写完的数据并没 有马上能够读到。

(3)最终一致性

最终一致性是弱一致性的一种特例。假如 ProcessA 首先写了一个值到存储系统，存储 系统保证如果在A、B、C 后续读取之前没有其他写操作更新同样的值，最终所有的读取操 作都会读取到 A 写入的最新值。此种情况下，如果没有失败发生，“不一致性窗口”的大 小依赖于以下几个因素：交互延迟、系统的负载以及复制技术中回滚的个数。最终一致性 方面最出名的系统可以说是域名系统 (DNS) 。 当更新一个域名的 IP 以后，根据配置策略 以及缓存控制策略的不同，最终所有的客户都会看到最新的值，但是有一个过程称为不一 致性窗口。

强一致性、最终一致性、弱一致性的比较如表3-2所示。

**表3-2** **强一致性、最终一致性、弱一致性的比较**

|  |  |  |  |
| --- | --- | --- | --- |
| **对** **比** **项** | **强** **一** **致** **性** | **最终一致性** | **弱** **一** **致** **性** |
| 场景定义 | 假设三个进程A、B、C相互独立，且都对存储系统进行读写操作 | | |
| 数据一致性 表现 | A写入数据到存储系统之后，存  储系统能够保证后续任何时刻  发起读操作的B、C可以读到A  写入的数据 | A写入数据到存储系统之  后，经过一定的时间，或  者在某个特定的操作之  后，B、C最终会读到A  写入的数据 | A写入数据到存储系统之  后，存储系统不能保证后  续任何时刻发起读操作  的B、C可以读到A写入 的数据 |
| 示例 | OLTP需要强一致性 | OLAP需要最终一致性 | 绝大多数应用不能够容 忍弱一致性 |

(4)因果一致性

如果Process A 通知 Process B它已经更新了数据，那么 Process B的后续读取操作则读 取A 写入的最新值，而与A 没有因果关系的C 则可以维持最终一致性。





第3章 分布系统设计的CAP 理论

(5)Read-your-writes 一致性

如果Process A写入了最新的值，那么 Process A的后续操作都会读取到最新值。但是 其他用户可能要过一会才可以看到。

(6)会话一致性

会话一致性要求客户端和存储系统交互的整个会话阶段保证 Read-your-writes一致性。 Hibernate的 session提供的一致性保证就属于这种一致性。

(7)单调读一致性

单调读一致性要求如果 Process A 已经读取了对象的某个值，那么后续操作将不会读取 到更早的值。

(8)单调写一致性

单调写一致性保证系统能序列化执行一个 Process中的所有写操作。

BASE 理论的精髓是选择分区容错性、可用性和最终一致性。

**3.5** **Web** **分布式系统设计**

设计大型 Web 系统时，需要注意的一些核心原则如下。

● 可用性；

● 性能；

● 可靠性；

● 可扩展；

● 易管理；

● 成本。

上面的这些原则为设计分布式 Web 架构提供了理论基础指导。但是，这些规则之间也 可能彼此相矛盾。例如，为了扩充存储容量，可以添加更多的服务器(可伸缩性),这个方 案是易管理，但成本高。因此，构建所有大型 Web应用程序都要考虑服务、冗余、分区和 故障处理能力，并且每个因素都会涉及选择和折中。

**3.5.1** **系统核心需求**

一些大型网站需要管理和传送大量的图片，需要构建一个具有低成本、高可用性和低 延时(快速检索)的架构。在图片系统中，用户可以上传图片到一个中央服务器，通过网 络连接或 API 对这些图片进行请求。简单地说，假设这个应用程序只包含两个核心部分： 上传(写)图片和检索图片。图片上传时最好能够做到高效，传输速度快，当一个 Web 页 面或其他应用程序向图片发出请求时，提供Web 服务或内容分发网络边缘服务器，由于可 以在许多地方存储内容，无论是在地理上还是物理上都更加接近用户，从而导致更快的性 能。该系统主要考虑下述几方面。

● 存储图片的数量无限制，所以存储系统应具备可伸缩性；

下载/请求需要做到低延迟；





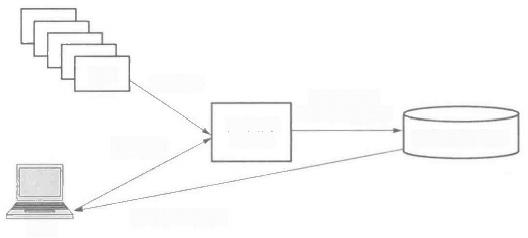
大数据基础与应用

● 用户上传一张图片，那么图片就应该始终存储在那里(图片数据的可靠性);

● 系统应该易于维护与管理；

● 由于图片托管没有太高的利润空间，所以系统需要较高的效益。 图片托管应用程序的功能如图3-11所示。

在这个例子中，系统必须快速、可靠和可扩展地存储数据。



上传文件

将图片文件

写入持久存储器

图片文件存储器

图片请求

回送请求的图片

用户

图片请求 处理器

图片

图3-11 图片托管系统的简化结构图

**3.5.2** **系统服务**

当构建可伸缩的分布系统时，所有上传和检索请求都在同一台服务器上处理。在这种 设计中，需要考虑下述问题。

写图片和读图片两个功能共享同一台服务器资源。 一般来说，下载速度是上传速度的 3 倍，通常文件可以从缓存中读取，而最终写入到磁盘中(也许在最终一致的情况下，可 以被多写几次)。即使是从缓存或者磁盘中读取，数据写入都比读慢。这种情况导致了对共 享同一台服务器资源存取速度的不匹配。

Web 服务器通常都有一个并发连接数上限值，读可以异步或利用其他性能优化方法， Web 服务可以快速切换读取客户端的更多的请求，超过每秒的最大连接数 (Apache 的最大 连接数设为500)。此外，写通常倾向于保持一个开放的链接进行持续上传，上传一个1MB 的文件花费的时间可能会超过1s, 所以，服务器只能同时满足500个写请求。

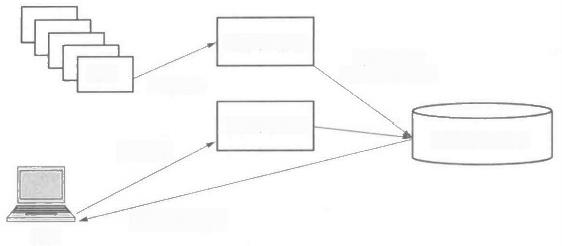
克服这种瓶颈的一种常用方法是将读过程和写过程分离，如图3-12所示。读过程和写 过程分离之后就可以对它们单独进行扩展。由于读操作比写操作多，更易于排除故障和解 决规模方面的问题，进而能够彼此独立解决问题，不必同时考虑写入和检索操作。

例 如 ，Flickr 解决这个读写问题的方法是通过分发用户跨越不同的碎片，每个碎片只能 处理一组用户，但是随着用户数的增加，更多的碎片也会相应地添加到群集里。在第一个 例子中，它更容易基于硬件的实际用量进行扩展，即在整个系统中的读写数量。Flickr 是基 于其用户群进行扩展。而前面的那个例子，任何一个中断或者问题都会降低整个系统功能 (例如，任何人都没办法执行写操作),而 Flickr 的一个中断只会影响到其所在碎片的用户 数。在第一个例子中，它更容易通过整个数据集进行操作，例如，更新写服务，包括新的 元数据或者通过所有的图片元数据进行搜索，而 Flickr 架构的每个碎片都需要被更新或搜 索，或者需要创建一个搜索服务来收集元数据。





第3章 分布系统设计的CAP 理论



图片写入服务

将图片文件

写入持久存储器

上传文件

图片取回服务

图片文件存储器

图片请求

回送请求的图片

用户

图片

图3-12 读取分离示意图

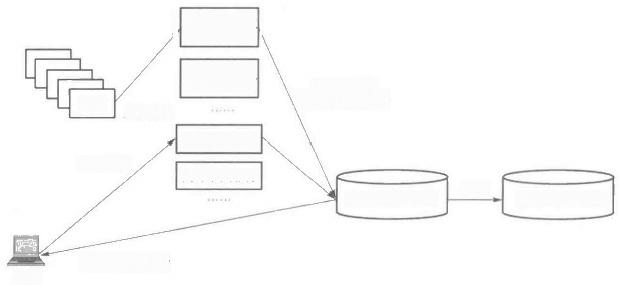
根据系统确定需求：大量的读写操作、级别并发、跨数据查询、范围、种类等，来选 择不同的基准、理解系统是如何出错的，对以后的故障发生情况做出初步的系统服务计划。

**3.5.3** **冗余**

为了正确处理错误， 一个Web架构的服务和数据必须具备适当的冗余。例如，如果只有 一个副本文件存储在一台单独的服务器上，那么这台服务器出现问题或丢失，该文件也随即 一起丢失。避免数据丢失的常用方法就是多创建几个文件或副本。这样的方法也同样适用于 服务器。如果一个应用程序的核心功能是应确保有多个副本或版本在同时运行，这样可以避 免单节点失败。在系统中创建冗余后，当系统发生危机时，如果需要，可以消除单点故障并 提供备份或备用功能。例如，有两个相同的服务示例在生产环境中运行，如果其中一个发生 故障，那么该系统容错转移至正常的副本上。容错转移可以自动发生也可以手动干预。

服务冗余的另一重要组成部分是创建一个无共享架构。在这种体系结构中，每个节点 都能相互独立运行，并且没有所谓的中央“大脑”管理状态或协调活动其他节点。这对系 统的可扩展帮助很大，因为新节点在没有特殊要求或知识的前提下被添加。然而，最重要 的是，这些系统是没有单点故障的，所以失败的弹性就更大。

例如，在图片服务器应用程序中，所有的图片在另一个硬件上都有冗余副本(理想情况下 是在不同的地理位置，避免数据中心发生火灾、地震等事故时导致数据不可用),如图3-13所示。



阳片写入服务

图片写入服务

上传文件

图片取回服务

图片请求

图片取回服务

图片文件存储器

回送请求的图片

用户

将图片文件 写入持久存储器

图片文件存储器

复制

图片

图3-13 图片托管应用程序冗余





**大数据基础与应用**

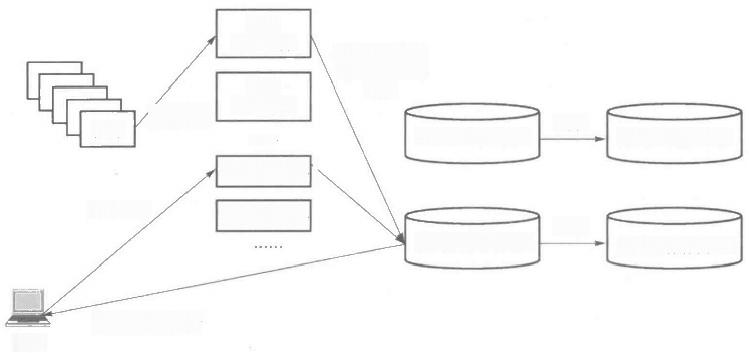
**3.5.4** **分区**

数据集有可能非常大，无法安装在一台服务器上。也有可能某操作需要太多的计算资 源、性能降低并且有必要增加容量。在这两种情况下，可有纵向扩展和横向扩展两种选择。

纵向扩展是在单个服务器上添加更多的资源。所以，对于一个非常大的数据集，这可 能意味着添加更多(或更大)的硬件设备，来使一台服务器能容下整个数据集，并拥有更 快的CPU 完成计算。在各种情况下，纵向扩展可以通过提升单个资源的处理能力来完成。

横向扩展是添加更多的节点，在大数据集下，这可能会使用多服务器来存储部分数据 集，对于计算资源，这意味着分割操作或跨节点加载。为了充分利用横向扩展，它应作为 一种内在的系统架构设计原则，否则修改或拆分操作将会非常麻烦。

当横向扩展时，最常见的做法是把服务进行分区或碎片化。分区可以被派发，这样每 个逻辑组的功能就是独立的，也可以通过地理界限或其他标准，如非付费与付费用户来完 成分区。这些方案的优点是它们会随着容量的增加提供一个服务或数据存储。在图片服务 器案例中，用来存储图片的单个文件服务器可能被多个文件服务器取代，每个里面都会包 含一套自己独特的图像，如图3-14所示。这种架构将允许系统来填充每一个文件服务器或 图片服务器，当磁盘填满时会添加额外的服务器。这样的设计需要一个命名方案，用来捆 绑图片文件名到其相应的服务器上。图像名字可以形成一个一致的哈希方案并映射到整个 服务器上。或者给每张图片分配一个增量ID, 当客户端对图片发出请求时，图片检索服务 只需要检索映射到每个服务器上(如索引)的ID。



图片

写入服务

将图片文件写

入持久存储器

图片

写入服务

复制

图片文件存储器1

图片取回服务

图片请求

复制

图片文件存储器2

回送请求的图片

用户

图片文件存储器2

图片文件存储器1

图片取回服务

上传文件

1或2

图片

…\*…

图3-14 图片托管应用程序冗余和分区

当然，跨越多个服务器对数据或功能进行分区还是有许多困难。其中的关键问题是数 据本地化。在分布式系统中，数据操作或计算点越接近，系统性能就会越好。因此，当数 据分散在多个服务器上时，如果数据不是在本地，那么就要迫使服务器通过网络来获取所 需的信息，这个获取的过程就会涉及成本。另一个问题是不一致。当有多个服务对一个共 享资源执行读写操作时，可能会有另一个服务器或数据存储参与进来。 一些数据需要更新，



**第3章** **分布系统设计的CAP理论**

但是读的优先级高于更新，在这种情况下，会导致数据不一致问题的发生。例如，在图片 托管方案中，有可能出现的不一致是：如果一个客户端发送更新“狗”图片请求，进行重 新命名，把 “Dog” 改成 “Gizmo”, 但同时，另一个客户端正在读这张图片。在这种情况 下，标题就是不清楚的。“Dog” 或 “Gizmo” 应该被第二个客户端接收。当然，在进行数 据分区时会产生一些困难，但是通过数据、负载、使用模式等，将把每个问题拆分到管理 群中。

以上所述是设计分布式系统需要考虑的核心要素。可用性、性能、可靠性、可扩展、 易管理、成本这几个原则非常重要，但在实际应用中可能会以牺牲某个原则来实现另外一 个原则，在这个过程中就要做好权衡工作。

**小结**

在大数据的采集、传输、存储与处理过程中，横向可扩展的分布系统是重要的环境， 而在横向可扩展的分布系统中， 一致性、可用性和分区容错性是系统的核心需求。CAP 定 理明确了实现分布式系统的局限性， 一个分布式系统不可能同时满足三个系统需求，最多 只能同时满足两个系统需求。在考虑满足系统需求时，要根据实际需要来选择关注点，进 而采用相应的策略。CAP 定理是设计分布系统的理论基础。本章介绍了CAP 定理的内容与 重要性，并从应用的角度出发，说明了CAP 定理的证明与应用。本章也介绍了BASE 模型 和Web 分布式系统设计的例子。



**第** **4** **章** **大数据网络空间**

本章主要内容



复杂网络概念与特征

复杂网络空间概述

复杂网络的特性

社会网络结构

社会网络理论

社会网络

社会计算

社会网络应用

社会网络分析概述

社会网络分析的原理

社会网络分析

社会网络分析的特征

社会网络分析的方法

用户隐私类型

身份隐私攻击与保护

社会网络中的隐私保护

用户关系的攻击及保护

万维网用户隐私保护

社会感知计算概念

社会感知计算

社会感知计算的内容

通信方式的演化

六度分隔理论

人类通信方式

150法则

唯象理论与唯象方法

社交网站作用

社交网站

Web网站

大数据网络空间

大数据网络空间是一个复杂的网络空间。在现实世界中，来自各领域的有意义的复杂 系统都可以映射成网络，大数据空间可以映射成大型复杂网络空间。



第 4 章 大数据网络空间

**4.1** **复杂网络空间概述**

形式最简单的网络是一系列通过线相连接的点，将点称为节点。因此，网络是一组由 边相连接的节点集合。现实世界中的系统是由个体元件以某种形式连接而成，例如因特网 是一系列通过数据连接的计算机构成，社会网络是一群通过熟人关系和社会交互互相关联 的人构成，食物链网络是一群通过扑食或被扑食相关联的动物构成，新陈代谢网络是一系 统通过化学变化相关联的代谢物构成。

网络数据科学应发现网络数据(信息)产生与传播的规律、网络信息涌现的内在机制 以及与其相关的社会学、心理学、经济学和信息科学的机理，利用这些机理研究互联网对 政治、经济、文化等各方面的影响。例如，因特网上的不同计算机之间的连接模式影响了

数据传输路径和传输效率。社会网的连接关系影响人们获 取信息、 建立主见、形成舆论，以及对一些不太明显的现 象产生影响。只有了解了这些网络结构，才能完全地理解

相应系统是如何工作的，因此，将复杂系统抽象成网络来 研究其模式完全必要、意义重大，参见图4-1。

**4.1.1** **复杂网络概念与特征**

复杂网络

复杂系统

大数据空间

图4-1

**1.** **复杂网络概念**

复杂网络是指具有自组织、自相似、吸引子、小世界、无标度中部分或全部性质的网 络，参见图4-2。

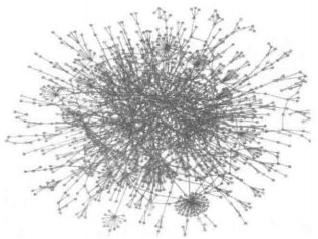


图4-2 复杂网络

**2.** **复杂网络特征**

复杂网络的复杂性主要表现在以下几个方面。

● 结构复杂：节点数目巨大，网络结构呈现多种不同特征。

● 网络进化：节点或连接随时可以产生与消失。例如，在万维网中的网页或链接随 时可能出现或断开，导致网络结构不断发生变化。





大数据基础与应用

● 连接多样性：节点之间的连接权重存在差异，且有可能存在方向性。

● 动力学复杂性：节点集可以属于非线性动力学系统，例如节点状态随时间发生复 杂变化。

● 节点多样性：节点可以代表任何事物。例如，人际关系构成的复杂网络节点代表 单独个体，万维网组成的复杂网络节点可以表示不同网页。

● 多重复杂性融合：以上多重复杂性相互影响，导致更为难以预料的结果。在设计 一个网络时，需要考虑此网络的进化过程，其进化过程决定了网络的拓扑结构。 当两个节点之间频繁进行能量传输时，它们之间的连接权重会随之增加，通过不 断的学习与记忆逐步改善网络性能。

复杂网络研究的内容主要包括：网络的几何性质、网络的形成机制、网络演化的统计 规律、网络上的模型性质以及网络的结构稳定性、网络的演化动力学机制等问题。其中在 自然科学领域，网络研究的基本测度包括：度及其分布特征、度的相关性、集聚程度及其 分布特征、最短距离及其分布特征、介数及其分布特征和连通集团的规模分布等。

**4.1.2** **复杂网络的特性**

复杂网络具有如下特性。

**1.** **小世界**

复杂网络尽管规模很大，但是任意两个节(顶)点间却有一条相当短的路径。表明相 互关系的数目可以很小，但却能够连接全世界。例如，在社会网络中，人与人相互认识的 关系很少，但是却可以找到距离遥远的无任何关系的其他人，呈现出世界变成一个地球村， 也就是说，变成一个小世界。

**2.** **集群**

集聚程度是指网络集群的程度，表现网络的内聚倾向，反映的是一个复杂网络中各集 聚的小网络分布和相互联系的状况。例如，在社会网络中，集群可以反映这个朋友圈与另 一个朋友圈的相互关系。

**3.** **幂律的度分布**

度指的是网络中某个顶(节)点(相当于一个个体)与其他顶点关系(用网络中的边 表达)的数量。度的相关性指顶点之间关系的联系紧密性，介数是一个重要的全局几何量。 顶点u 的介数含义为网络中所有的最短路径之中，经过u 的数量。它反映了顶点u (即网 络中有关联的个体)的影响力。

网络在表征系统元件的交互连接模式上是一种非常有力的工具，网络是复杂系统的最 简单的一种表征形式。网络将复杂系统简化为仅仅体现基本的连接模式和极少其他信息的 抽象结构。网络中的节点和边可以标识为其他额外信息，如名字或强度等，这样可以更多 体现系统的细节信息。利用网络表征，将来可以把生物学、社会学、工程学、信息学等不 同领域的复杂系统抽象为复杂网络。简而言之，复杂网络即呈现高度复杂性的网络，而针





**第** **4** **章** **大数据网络空间**

对复杂网络研究出的各种算法和工具，理论上可以适用于能用复杂网络表征的任何复杂系 统。复杂网络研究有着广泛的应用前景。可以断言，20世纪是量子力学的世纪，而21 世 纪，网络理论将成为量子力学的继承者。

**4.2** **社** **会** **网** **络**

社会媒体是具有广泛用户的在线交互媒体，允许用户在线发布和传播信息，相互沟通 与协作，组成虚拟网络社区。社会媒体具有媒体属性和社交功能，其基础是社会网络。目 前最受关注的社会媒体是社交网站 (Social Networking Services,SNS) 和微博，国外以 Facebook 和 Twitter 为代表，国内以人人网、新浪微博、腾讯微博为代表，典型的社会媒体 还包括论坛、博客、维基百科和视频网站等。社会网络已经逐渐成为人们交流、传播、分 享信息的主要媒介。社会网络产生的理念来源于六度分隔理论和150法则，社交网络构成 了大数据的重要生态环境。社会网络的研究意义深远。

社会网络出现于20世纪30年代，社会结构的概念不断地深化，形成了一套系统的理 论、方法和技术，并成为一种重要的社会结构研究范式。互联网是机器的互联，万维网是 信息的互联，物联网是物的互联，社会网是人的互联。

社会网络的概念首先由英国著名人类学家 R.布朗提出，社会网络是由某些个体间的社 会关系构成的相对稳定的系统，把网络视为联结行动者的一系列社会联系或社会关系，其 相对稳定的模式构成了社会结构。随着应用领域的不断扩展，现在的社会网络已经超越了 人际关系的范畴，网络的行动者可以是个人，也可以是集合单位，如家庭、部门、组织。 网络成员占有各种不同的稀缺性资源，关系的数量、方向、密度、力量和行动者在网络中 的位置等因素，影响资源流动的方式和效率。

**4.2.1** **社会网络结构**

社会网络由社会关系所构成，社会网络表示了行动者之间的社会结构关系。构成社会 网络的主要要素如下。

● 行动者：行动者可以是具体的个人，还可以是一个群体或集体性的社会单位。每 个行动者在网络中的位置被称为节点。

● 关系：将行动者之间相互的关联称为关系。主要的关系形式是亲属关系、合作关 系、交换关系、对抗关系等。

● 二人组：二人组是社会网络中最基本的形式，由两个行动者所构成的二人组关系 是分析各种关系的基础。

● 三人组：三人组关系是指由三个行动者所构成的关系。

● 子群：子群关系是指行动者全集的子集。

● 群体：其关系得到测量的所有行动者的集合。





**大数据基础与应用**

**4.2.2** **社会网络理论**

社会网络理论由关系要素和结构要素组成。关系要素关注行动者之间的社会性关系， 通过社会联结的密度、强度、对称性、规模等来说明特定的行为和过程。结构要素则关注 网络参与者在网络中所处的位置，研究两个或两个以上的行动者和第三方之间的关系所折 射出来的社会结构，以及这种结构的形成和演进模式。这两类要素都对知识和信息的流动 有着重要的影响。更具体地说，联结强度、社会资本、结构洞是社会网络理论的核心内容。

**1.** **联结强度**

社会网络的节点间的联结分为强联结与弱联结，可以从互动的频率、感情力量、亲密 程度和互惠交换四个维度来进行区分。强联结与弱联结又分别称为强关系与弱关系。强关 系是指在性别、年龄、教育程度、职业身份、收入水平等社会经济特征相似的个体之间的 关系，而弱关系则是在社会经济特征不同的个体之间的关系。群体内部相似性较高的个体 所了解的事物、事件经常是相同的，所以通过强关系获得的资源常是冗余的。而弱关系是 在群体之间发生的，跨越了不同的信息源，能够充当信息桥的作用，可将其他群体的信息、 资源带给本不属于该群体的某个个体。强关系维系着群体、组织内部的关系，弱关系在群 体、组织之间建立了纽带联系。通过强关系获得的信息往往重复性很高，而弱关系比强关 系更能跨越其社会界限去获得信息和其他资源。社会网络的节点间的联结是社会网络分析 的最基本分析单位。

弱联结是获取无冗余的新知识的重要通道，但是，资源不一定总能在弱联结中获取， 强联结是个人与外界发生联系的基础与出发点。网络中发生的知识的流通往往发生于强联 结之间。强联结包含某种信任、合作与稳定，能够传递高质量的、复杂的或隐性的知识。

**2.** **社会资本**

法国社会学家布迪厄 (Bourdieu) 提出了社会资本概念。社会资本指个人所拥有的， 但表现为社会结构资源的资本财产。它们由构成社会结构的要素组成，主要存在于社会团 体和社会关系网之中。个人参加的社会团体越多，其社会资本越雄厚。个人的社会网络规 模越大、异质性越强，其社会资本越丰富。社会资本越多，摄取资源的能力越强。不仅个 人具有社会资本，企业也有企业社会资本，通过联结获取稀缺资源的能力就是企业的社会 资本。由于社会资本代表了一个组织或个体的社会关系，因此，在一个网络中， 一个组织 或个体的社会资本数量决定了在网络结构中的地位。

**3.** **结构洞**

无论是个人还是组织在社会网络中表现为两种形式， 一种形式是在网络中的任何主体 与其他主体都存在联系，不存在联系断开现象，从整个网络来看，这就是无洞结构。由于 无洞结构要求主体之间全联结，所以只在小群体中存在。另一种形式是社会网络中的某个 或某些个体与其他个体不发生直接联系，无直接联系或联系中断的现象，从整个网络来看， 无直接联系就是出现了洞。





**第4章** **大数据网络空间**

可以看出， 一个结构洞表明了两个行动者之间的非冗余的联系。例如，对于三个行动 者 A 、B 、C来说，如果A 和 B 有联系，A 与 C 有联系，但是B 和 C 之间不存在联系，那 么 B 和 C 之间就相当于存在一个洞。这种结构的 A 、B 、C 之间关系就是一个结构洞。A 是结构洞的中间人。结构洞能够为中间人获取信息利益和控制利益提供机会，进而比网络 中其他位置上的成员更具有竞争优势。

由于结构洞之内填充的是弱联结，因而结构洞可以看作是对联结强弱观点的深化与系 统化。另外，结构洞与社会资本有关，主体拥有的结构洞越多，具有的社会资本越多。

**4.2.3** **社会计算**

社会大数据催生了社会计算学科的发展，利用社会计算可以架起社会科学和计算科学 之间的桥梁，可以从基础理论、实验手段以及应用等层面进行社会科学与计算科学的交叉 性研究。主要研究内容是应用网络数据来研究群体社会行为及其演化规律，这也标志着计 算科学和社会科学的交叉融合正成为新的研究热点。

社会计算主要集中在计算社会科学、社会计算应用和群体智慧方向的研究。计算社会 科学研究利用计算技术揭示社会运行的规律与趋势，社会计算应用研究利用计算系统帮助 人们沟通与协作，群体智慧方向的研究在机器的辅助下以人类群体协作的方式解决问题。

计算社会科学和社会计算应用都是面向社会的科学技术，计算社会科学基于科学层面， 社会计算基于技术层面，群体智慧基于社会层面，它是一种全新的基于人脑互联的计算模 式。社会计算是人类认识社会本质规律的强大研究工具。社会计算处理的对象，例如社交 网站与社会媒体等，其背后是一个巨大的社会网络。社会网络是一个复杂系统，需要利用 复杂系统研究成果应用于社会网络。

**1.** **个体与群体的社会建模**

为了构建社会个体或群体的行为、进行认知和心理模型以及对社会群体的行为特点进 行分析，应该对个体与群体的社会建模、对社区结构、交互模式、个体间的社会关系等的 建模。许多社会科学的理论模型都与个体和群体的社会建模相关。社会心理学揭示了社会 认知与心理的形成机制及其发展的基本规律，社会动力学研究人类社会发展的动态过程及 其演化规律，社会物理学研究社会稳定的机理以及人类行为模式与社会稳定的关系等。基 于计算的社会个体与群体的研究大多基于文本数据。社会网络是反映个体间社会交往与互 动关系的主要手段，通过网络节点间的链接关系来发现并识别潜在的社会群体。

**2.** **社会文化建模与分析**

社会文化建模与分析主要包括基于社会文化因素建模、基于智能体的人工社会建模、 计算实验分析、人工社会系统与计算实验平台设计等。利用计算技术来研究文化冲突和变 迁，分析不同文化背景的国家或组织的决策过程，探寻其行为所依赖的社会文化因素，已 成为社会计算建模的重要研究方向。由于社会事件的出现往往具有突发性和不可重复性， 因此，采用传统方法对其演化过程进行实验分析和评估将是一项十分困难的研究工作。



**大数据基础与应用**

**3.** **社会交互及其规律分析**

社会交互及其规律分析主要是针对人群交互行为的特点及社会事件演化规律进行分 析，也包括社会网络结构、信息扩散和影响、复杂网络与网络动态性、群体交互和协作等 的分析。计算社会学认为网络上的大量信息，如博客、论坛、聊天、消费记录、电子邮件 等，都是现实社会的人或组织行为在网络空间的映射。利用网络数据可分析个人和群体的 行为模式，从而深化对生活、组织和社会的理解。计算社会学的研究涉及人们的交互方式、 社会群体网络的形态及其演化规律等问题。

**4.** **社会数据感知与知识发现**

社会数据的获取和规律性知识的挖掘的主要内容包括社会学习、社会媒体分析与信息 挖掘、情感及观点挖掘、行为识别和预测等。社会数据的主要形式包括文本、图像、音频、 视频等，这些数据除了来自网络媒体信息(包括博客、论坛、新闻网站等),还来自专用网 络、传统媒体和应用部门的内部数据等。通过构建社会传感网络，可以有效利用数据源所 隐含的社会化信息的结构特征。通过对重要节点信息的动态监控，实现对社会数据的全方 位、分层次感知。基于社会数据的知识发现包括对社会行为和心理的分析与挖掘。

**5.** **决策支持及应用**

在社会经济与安全等领域应用社会计算，可以对管理者和社会提供决策支持、应急预 警、政策评估和建议等。因为网络社会媒体能够充分体现人们的价值取向和真实意愿，所 以比传统媒体的反应更灵敏、准确。

**4.2.4** **社会网络应用**

知识获取、知识传递、知识共享与知识创新是重要的活动。知识活动发生在社会关系 网络中，必然受到社会网络特性的制约和影响。

**1.** **知识获取分析**

知识获取是知识活动的主体从周边环境中寻找知识来源，以获得所需要的知识或信息 的过程。在一个组织内，各行为主体(组织成员)频繁接触而组成的关联网络是知识获取 的主要来源。相似度高的个体所了解的事物和社会经历在很大程度上相同，因此在强联结 的同质群体内，知识以固有形态存在，个体难以获得新的知识。弱联结所联系的是两个社 会经济特征不同的个体，它们嵌入在不同的社会网络中，拥有不同的信息源。弱联结可以 跨越不同的信息源，起到沟通和连接网络中不同群体(如不同职能部门)的作用。因此， 弱联结降低了获取知识的难度和成本。弱联结带来的新知识有利于企业迅速掌握新技术， 从而提高竞争力。下述两种网络位置有利于企业层面的知识获取。

(1)结构空洞位置

一个网络中最有竞争优势的位置是在结构空洞上。在结构空洞的企业能有机会接触到 两种异质的信息流，获得无冗余信息。同时作为信息流动的必经节点，具有相对控制优势，





第4章 大数据网络空间

所以处在结构空洞位置的企业具有信息优势。

(2)网络中的核心位置

如果企业处在网络中的核心位置，那么核心企业掌握大量网络成员生存的必要资源， 其他成员对核心企业高度依赖，使核心企业被看作期望的潜在合作者，能够参与一系列的 重要联结，从而在网络中占据战略性位置。核心企业因为拥有较多的联结关系和社会资本， 对于网络中的资源流动具有支配权，能够更快地获得各方面的知识和技术支持。

**2.** **知识类型与传递**

不同的社会网络关系，适合传递不同的知识类型。

(1)知识类型

可以将知识划分为显性知识和隐性知识两种类别。

① 显性知识是可以表达的、有物质载体的、可确知的。事实知识和原理知识基本属于 显性知识。

② 隐性知识是个人或组织经过长期积累而拥有的知识，不易用言语文字表达，传播起 来非常困难。隐性知识所对应的是技能知识和人力知识。

(2)弱联结有利于简单信息的传递

弱联结有利于简单信息的传递，促进事实知识的分享。IT 技术的发展给人们提供了新 的互动环境和交流空间，使现实人际交往逐步向网络世界延伸。由于个人具有更多的弱联 结关系，所以弱联结在显性知识传递中的作用更加突出。

(3)强联结有利于隐性知识的传递

当组织间具有较强的社会联结、信任关系，具有相同的价值观、规范时，组织间的知 识传递会更有效率。强联结能够有效促进技能知识和隐性知识的传递和共享。

① 强联结可以保持人与人之间的密切接触，空间位置的接近有利于隐性知识的流动。 ② 强联结能够增加社会资本，促进隐性知识的分享和扩散。

(4)知识创新

知识在个体的大脑中产生，然后在脑与脑之间传递，从而形成更大量、更综合、更系 统的知识。在人与人的社会互动过程中，知识创新时有发生。在社会网络中的成员所拥有 的知识资源既有交集，又有补集。网络中容纳了不同个性、观点、理念，通过知识的搜寻、 传递、共享，这些不同的思想和智慧发生冲突，或达成共识，并在此过程中不断升华和超 越，最终产生新知识。

**4.3** **社会网络分析**

社会网络分析起源于适应性网络，通过研究社会网络关系，可以把个体间关系、微观 网络与大规模的社会系统宏观结构联系起来，利用数学定量分析方法，逐步发展成一个新 的分支领域。





**大数据基础与应用**

**4.3.1** **社会网络分析概述**

**1.** **社会网络分析的内容**

社会网络分析是通过对于网络中关系的分析，研究网络的结构及属性特征，包括网 络中的个体属性及网络的整体属性。网络个体属性分析的主要内容包括：点度中心度与 接近中心度等。网络的整体属性分析主要包括小世界效应、小团体研究与凝聚子群等。 社会网络分析是对社会网络的关系结构及其属性加以分析的一套规范和方法，又被称为 结构分析法，它主要分析的对象是不同单位(个体、群体或社会)所构成的社会关系的 结构及其属性。

社会网络分析主要包括对人物节点的分析、社会关系的分析、社会群体的分析和社会 网络拓扑结构分析等。

**2.** **社会网络分析的思维方式**

在社会环境中，可以将人的相互作用表达为关系的模式或规则，而这种关系的模式反 映了社会结构，这种结构的量化分析正是社会网络分析的出发点。社会网络分析是一种基 于关系论的思维方式，可以用来解释一些社会学、经济学和管理学等领域的问题。例如， 社会网络分析在职业流动、城市化对个体的影响、世界政治和经济体系、国际贸易等领域 已有广泛应用。

**3.** **社会网络分析的意义**

社会科学研究的对象应是社会结构，而不是个体。通过研究网络关系，有助于把个体 间关系、微观网络与大规模的社会系统宏观结构结合起来。社会网络分析可以为社会结构 的新理论奠定基础。

对社会现象的研究存在个体主义方法论与整体主义方法论两种对立的方法论。前者强 调个体行动及其意义，认为对社会的研究可以转换为对个体行动的研究。整体主义方法论 重视对社会结构的研究。在社会学中，社会结构是在各不相同的层次上体现的。它既可用 以说明微观的社会互动关系模式，也可说明宏观的社会关系模式。也就是说，从社会角色 到整个社会，都存在着结构关系，社会结构有多重含义。

**4.3.2** **社会网络分析的原理**

社会网络分析是研究社会结构的一种基本方法，其基本原理如下。

● 关系经常是不对称的相互作用，在内容和强度上有所不同。

● 关系间接或直接地把网络成员连接在一起。

● 社会结构产生了非随机的网络，因而形成了网络群、网络界限和交叉关联。

● 交叉关联把网络群以及个体联系在一起。

● 不对称的关系和复杂网络使稀缺资源的分配不平等。

第4章 大数据网络空间

**●** **网络产生了以获取稀缺资源为目的的合作和竞争。**

**4.3.3** **社会网络分析的特征**

社会网络分析的特征如下。

● 社会网络分析是根据结构对行动的制约来解释人们的行为。

● 社会网络分析关注不同单位之间的关系分析。

● 社会网络分析主要研究的是由多维因素所构成的关系形式。

● 社会网络分析将结构作为网络间的网络。

● 社会网络分析直接涉及的是具有一定的社会结构的关系性质。

综上所述，根据社会网络分析的基本思想，即行动者的任何行动都不是孤立的，而是 相互关联的。他们之间所形成的关系是信息和资源传递的渠道，网络关系结构也决定了行 动机会及结果。

**4.3.4** **社会网络分析的方法**

可以从中心性分析、凝聚子群分析、核心—边缘结构分析以及结构对等性分析等多个 不同出发点对社会网络进行分析，这里仅介绍3种常用方法。

**1.** **中心性分析**

在中心性分析方法中，个人或组织在其社会网络中的权力或地位代表了个体的中心度， 它表明了该节点在网络中的重要程度。因此一个网络中有N 个行动者(节点),就有N 个 个体的中心度。除了个体的中心度外，还引入整个网络的集中趋势，简称为中心势。网络 中心势反映的是整个网络中各个点的差异程度，因此一个网络只有一个中心势。根据计算 方法的不同，中心度和中心势都可以分为3种：点度中心度、点度中心势，中间中心度、 中间中心势，接近中心度、接近中心势。

(1)点度中心性

可以用网络中与该点有联系的点的数目来表示社会网络中一个点的点度中心性。如果 一个行动者与其他行动者之间存在直接联系，那么该行动者就居于中心地位，在该网络中 拥有较大的权力。网络中心势表示点的集中趋势，计算方法如下：首先找到图中的最大中 心度，然后计算该值与任何其他点的中心度的差，从而得出多个差值，再计算这些差值的 总和，最后用这个总和除以各个差值总和的最大可能值。可以看出，无论是点度中心势， 还是点度中心度都反映了点度中心性。

(2)中间中心性

在网络中，如果一个行动者处于许多其他两点之间的路径上，可以认为该行动者位于 重要地位，具有控制其他两个行动者之间的交往能力。根据这种思想来反映行动者个体中 心度的是中间中心度，它表明的是行动者对资源控制的程度。 一个行动者在网络中占据这 样的位置越多，就越代表它具有很高的中间中心性，就有越多的行动者需要通过它才能发 生联系。中间中心势也是分析网络整体结构的一个指数，其含义是网络中中间中心性最高





**大数据基础与应用**

的节点与其他节点的中间中心性的差距。该节点与别的节点的差距越大，则网络的中间中 心势越高，表示该网络中的节点可能分为多个小团体，而且过于依赖某一个节点传递关系， 该节点在网络中处于极其重要的地位。

(3)接近中心性

点度中心度描述的是局部的中心指数，表明的是网络中行动者与他人联系的多少，没 有考虑到行动者能否控制他人。而中间中心度描述的是一个行动者控制他人行动的能力。 这种能力可用接近中心性来描述。在计算接近中心度的时候，关注的是捷径，而不是直接 关系。如果一个点通过比较短的路径与许多其他点相连，我们就说该点具有较高的接近中 心性。对社会网络来说，接近中心势越高，表明网络中节点的差异性越大；反之，则表明 网络中节点间的差异越小。

**2.** **凝聚子群分析**

当网络中某些行动者之间的关系特别紧密，以至于结合成一个次级团体时，这样的团 体在社会网络分析中被称为凝聚子群。分析网络中存在多少个这样的子群，子群内部成员 关系的特点，子群之间关系特点， 一个子群的成员与另一个子群成员之间的关系特点等就 是凝聚子群分析。由于凝聚子群成员之间的关系十分紧密，因此又将凝聚子群分析称为小 团体分析。

**3.** **核心—边缘结构分析**

核心—边缘结构分析的目的是研究社会网络中哪些节点处于核心地位，哪些节点处于 边缘地位。核心边缘结构分析具有较广的应用性，可用于分析精英网络、科学引文关系网 络以及组织关系网络等多种社会现象中的核心一边缘结构。

**4.4** **社会网络中的隐私保护**

社会网络已经延伸到虚拟网络环境中，通过基于网络的互动服务，如聊天、实时消息、 文件分享、博客、讨论组等方式，用户可以相互交流和分享信息。互联网所具有的跨越时 空的特点促使虚拟的社会网络迅猛发展，对人们的生活和工作产生了深刻的影响。社会网 络已成为一种新型的协同工作方式，完善和改进了传统的沟通和交互机制。在社会网络中， 存放了大量用户隐私数据， 一旦被恶意攻击者窃取和使用，将为人们的生活、学习和工作 带来不良后果，甚至造成财产和声誉损失。随着社会网络的普及和发展，社会网络聚集了 众多用户，有效保护用户隐私将对社会安定和发展具有深远影响。

**4.4.1** **用户隐私类型**

隐私权是指自然人享有的私人生活安宁与私人信息秘密依法受到保护、不被他人非法 侵扰、知悉、收集、利用和公开的基本权利。网络隐私权是隐私权在网络空间中的体现， 是人们在网上享有对个人数据、私人信息、个人领域有知情权、选择权、合理的访问控制



第4章 大数据网络空间

权，以及保证其安全性和请求司法救助的权利。社会网络中的用户隐私表述如下。

**1.** **身份信息**

身份信息(即属性信息)是指唯一标识现实生活中特定个体的数据，如用户的姓名、 年龄、性别、身份证号码、民族和职业等，通常列为保密信息。

**2.** **私人数据**

私人数据是指用户在社会网络中发布的有关个人行为或价值取向等方面的信息， 一般 为受限访问的信息。信息拥有者可以通过网站提供的权限管理方法和工具，限定访问群体 和操作。

**3.** **用户间关系**

用户间关系指在社会网络中用户之间的关系，如是否存在联系、紧密程度如何等。这 种关系可能存在于同一个社交服务网站，也可能在多个不同类型的社交服务网站中。社会 网络一般是由节点和联系两大部分构成。其中，节点表示社会网络中的行为主体，即用户 或组织。边表示节点之间的联系，它是基于用户之间的某些特定关系而建立，如共同兴趣 爱好或贸易活动等。

恶意攻击者企图利用用户的个人信息谋取利益，网络技术和信息通信技术的发展在某 种程度上为侵害网络隐私权的行为提供了便利手段。在社会网络活动中，用户通常需要提 供部分个人信息，并认为泄露有限的信息是无关紧要的，但是攻击者通过采用信息检索或 数据挖掘等技术，找出信息间的联系，从而获得用户的大量完整隐私信息。攻击者也能在 未经用户允许的情况下私自传播、滥用和篡改用户信息，从而损害用户的利益。

**4.4.2** **身份隐私攻击与保护**

在社会网络中，用户使用身份信息进行注册，并采用匿名化技术保护这些信息，即隐 藏发布数据中能够标识用户的相关信息。去匿名化攻击是指恶意攻击者利用已经掌握的背 景知识，如网络拓扑结构、节点之间的联系等，结合用户的公开信息进行去匿名化推理，

以获取用户身份信息。

为了识别出潜在的去匿名化攻击，需提出一种分析和评估匿名化程度的度量方法，如 采用信息熵度量的方法，从而客观地比较不同匿名化系统或策略的效率，为制定出更加可 靠和高效的匿名化策略提供依据。将个人隐私分类，分别量化隐私信息泄露造成的风险， 并制订出不同的解决方案。

**4.4.3** **用户关系的攻击及保护**

面向用户关系的攻击是恶意攻击者利用所掌握的背景知识，借助关系挖掘技术、对社 会网络用户潜在的关联信息进行挖掘，从而获得用户之间的关系。针对这种攻击，典型的 防范方法是随机化图修改法，将原有社会网络图，随机删除K 个真实的边，然后添加K 个





**大数据基础与应用**

虚假的边，保证了原有图中节点的总数不变，从而保护节点间的敏感连接。

**4.4.4** **万维网用户隐私保护**

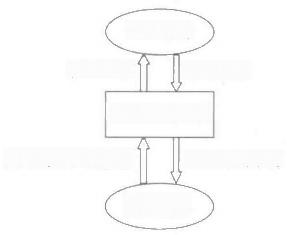
万维网联盟制定了隐私偏好规范，将其作为万维网站点定义相关隐私策略的方法，为 万维网浏览器自动读取和处理策略以及解决隐私保护问题提供了标准。隐私偏好规范由词 汇表和数据元素构成，定义了描述数据处理方式的标准词汇以及用于描述所收集信息种类 的基本数据模式。与现实中人们对隐私偏好的细节描述不同，隐私偏好规范描述的隐私策 略由一系列多项选择组成，这种标准格式适于系统自动处理。该规范还包含了用于请求和 传输隐私偏好规范策略的协议。万维网用户可以通过隐私偏好规范技术从服务器获取隐私 策略，并自动完成隐私策略匹配工作。隐私偏好规范技术为用户提供了描述个人隐私保护 策略的方法以及实施细粒度的访问控制机制，从而增强了用户对个人隐私信息的控制权。

**4.5** **社会感知计算**

**4.5.1** **社会感知计算概念**

社会感知计算是指通过人类社会生活空间日渐部署的大规模多种类传感设备，实时感 知识别社会个体的行为，分析挖掘群体社会交互特征和规律，辅助个体社会行为，支持社 群的互动、沟通和协作，从而高效地支持社会目标的实现。社会感知计算具有两层意义，

首先感知物理世界，然后觉察并做出响应。与基于单一万维网数据或用户调查数据的社会 计算或社会网络分析不同，社会感知计算强调利用先进计算机科学技术感知现实世界个体 行为和群体交互，理解人类社会活动模式，并为个体和群体交互提供智能辅助和支持，如 图4-3所示。社会感知计算借助普适环境新型智能设备和技术可以感知物理世界实时、连 续的现场数据，经过分析和处理，获得个体和群体的交互信息，为宏观社会提供决策依据。 反过来，社会感知计算从宏观社会接受理论指导，通过普适环境的大量智能设备，直接反 馈并作用于物理世界，辅助支持人类的社会活动。因此，社会感知计算是连接宏观社会和 物理世界的桥梁。

宏观社会

决策根据 理论指导

社会感知计算

实时连续现场数据 直接反馈作用

物理世界

图4-3 连接宏观社会和物理世界的社会感知计算





第4章 大数据网络空间

**4.5.2** **社会感知计算的内容**

社会感知计算的研究内容主要包含以下五个方面。

**1.** **现实世界实时数据感知**

利用现实物理世界部署的大规模多种类传感设备实时感知物理世界社会个体的活动原 始数据。研究不同模态数据的语义表示和关联。研究大规模感知数据的汇聚、融合和存储。

**2.** **人类行为与交互分析**

研究个体行为识别，分析其行为特性和规律。在个体行为、移动性和社会交互数据的 基础上，利用社会网络分析、机器学习、数据挖掘等方法，研究分析群体社会交互，如群 体形成、组织结构、群体活动、交流模式及其动态演化。

**3.** **社会交互高效支持**

在了解个体社会行为、群体社会交互状态和规律的前提下，研究社会交互高效智能支 持策略和机制。具体技术包括个性化推荐、社会交互状态可视化、群体协作支持、智能决 策等。

**4.** **社会感知计算软件框架和方法学**

研究社会感知计算基础软件框架，为系统的开发、测试和部署提供支撑。研究社会感 知计算方法论，制定系统设计原则，从系统的角度指导模型和算法的设计。研究社会感知 计算系统判断和评价的标准和方法。

**5.** **社会感知计算应用**

研究社会感知计算理论和技术在不同领域的应用。社会感知计算可以应用在很多重要 领域，如健康卫生(传染病防范)、公共安全(突发事件预警)、大规模系统工程(群体协 作支持)、智能交通管理(道路交通协同监测)、城市规划与发展(人口、资源、环境预测 与规划)等。

科学家对真理的不懈探索，推动了技术进步和社会发展，反过来新的技术和应用又影 响并改变了科学家研究的内容和方法。社会感知计算正是在普适计算技术的发展下问世， 由此产生新的科学问题，新的研究方法、工具和数据，使得感知物理世界、分析社会交互、 支持宏观社会成为可能。社会感知计算为计算机科学的研究与发展带来机遇，通过计算 机科学与人文社会科学、认知科学、社会心理学的交叉融合，感知并揭示人类社会内在 活动规律(与自然规律相对应),体现了重大基础和原始创新研究。开展社会感知计算研 究，形成具有独立自主知识产权的理论体系和支撑技术，提供社会性辅助和支持，提高 社会活动的效益和水平，增强社会凝聚力并实现和谐社会，将具有重大的科学意义和社 会价值。





**大数据基础与应用**

**4.6** **人类通信方式**

Internet 的出现，尤其是云计算与移动计算，促进了人类通信方式的发展。

**4.6.1** **通信方式的演化**

从通信方式的演化进程看出，最初的通信方式都是近距离的，包括一对一面谈、 一对 多演讲和多对多的会议。但是人们不满足于当面交流，希望能够突破时空。人们经过努力 实现了从近距离通信到远距离通信的空间突破，实现了从同步通信到异步通信的时间超越。 分别实现了电话、即时通信、广播、电视、视频会议远距离的一对一、一对多和多对多的 同步通信，实现了信件(无论是传统的信件还是电子邮件)和报纸、网站远距离一对一和 一对多的异步通信。但是，直到 Web 2.0 网站出现之后，才产生了空前的信息传播效果， 实现了多对多、远距离、最复杂的异步通信形式。人类通信方式的演化如表4-1 所示。

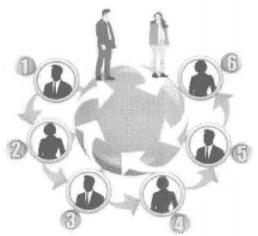
**表4-1** **人类通信方式的演化**

|  |  |  |  |
| --- | --- | --- | --- |
| **距离**  **方式** | **近距离(同步)** | **远距离(同步)** | **远距离(异步)** |
| 一对一 | 面谈 | 电话、即时通信 | 信件 |
| 一对多 | 演讲 | 广播、电视 | 报纸、网站 |
| 多对多 | 现场会议 | 视频会议 | Web 2.0社会媒体 |

在各种社会媒体中，与语言处理关系最密切的是微博。微博以简短的文字记录用户 的所闻所见。用户之间通过关系建立起人与人之间的关注网络， 一个用户发出的微博可 以快速传播。这些特点使微博迅速成为一种新的社会媒体，并且其用户群不断得到扩大。 Web 2.0 网站与 Web1.0 网站的比较如表4-2所示。

**表4-2** **Web2.0网站与Web 1.0 网站的比较**

|  |  |  |
| --- | --- | --- |
| **网络关系、结构和信息内容** | **Web** **1.0网站** | **Web** **2.0网站** |
| 信息源 | 网站的编辑 | 广大用户 |
| 节点 | 静态网页 | 人 |
| 节点分析 | 文本分析 | 用户兴趣分析 |
| 边 | 超链接 | 关注关系边是信息转发通道 |
| 边的分析 | 链接分析 | 人物关系分析、社交圈分析 |
| 节点重要性计算 | 网页重要性计算 | 意见领袖发现 |
| 节点间互动 | 转载 | 转发、评论 |
| 网络结构 | 变动慢 | 变动较快 |





第4章 大数据网络空间

从表4-2 中可以看出，最根本的变化是网络的节点从网页变成了人，这个根本性的变 化导致网络关系、网络结构以及信息内容等各方面发生变化。与以往网络媒体不同，社会 媒体后面的社会网络，以及借助社会网络而产生的社会话题传播的速度很快。要在社会媒 体环境下开展语言计算的工作，必须充分利用社会网络的信息和话题传播的模式，在此基 础上运用自然语言处理技术实现对社会媒体中人物和信息的更深层次的理解。

**4.6.2** **六度分隔理论**

1967年，美国著名社会心理学家、哈佛大学社会心理学教授 Stanley Milgram(1933— 1984)提出了六度分隔理论假说。科学发现一般经过两大步骤，第一步是提出假设，第二步是 给出证明，这同其他任何科学发现是一样的，提出合理假设并不比给出证明简单、容易。在许 多时候，显得更为重要，因为它指给人们的努力方向。赫赫有名的哥德巴赫猜想便是一例。

**1.** **六度分隔理论描述**

六度分隔理论是小世界现象，又称小世界效应，简单地描述为：在人际脉络中，要结 识任何一位陌生的朋友，这中间最多只要通过六个朋友就能达到目的。想认识一个人，托 朋友找认识他的人，总计不会超过六个人就可完成。也就是说，最多通过六个人就能够认 识任何一个陌生人，其示意图如图4-4所示。

20世纪60年代，米尔格兰姆 (Stanley Milgram) 设计了一个连锁信件实验，将一套连 锁信件随机发送给居住在内布拉斯加州奥马哈的160个人，信中放了一个波士顿股票经纪 人的名字，信中要求每个收信人将这套信寄给自己认为是比较接近那个股票经纪人的朋友。 朋友收信后照此办理。最终，大部分信在经过五六个步骤后都抵达了该股票经纪人。

按照六度分隔理论，每个个体的社交圈不断扩大，最后成为一个大型社会交际网。根 据这种理论，创立了面向社交网络的互联网服务，可以通过熟人的熟人来进行基于网络的 社交拓展。这个玄妙过程引起了数学家、物理学家和计算机科学家的兴趣，并进行研究后 发现，许多其他的网络也有相似的结构。熟人的熟人只是社交拓展的一种方式，社交网络 服务的含义已经远不只熟人的熟人这个层面。例如，根据相同话题进行凝聚、根据学习经 历进行凝聚、根据周末出游的相同地点进行凝聚等，都可被纳入社交网络服务的内容。如 图4-5所示是六度分隔理论的抽象表示，其中节点表示人，连线表示联系。

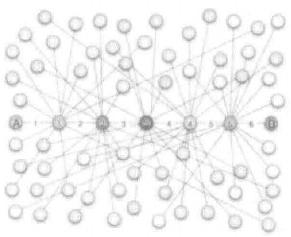


图4-4 六度分隔理论示意图

图4-5 六度分隔理论



**大数据基础与应用**

在现实社会中，社会交际是通过人与人之间的介绍、握手来形成一个朋友圈与联系圈， 每个人不需要直接认识所有人，只需要通过他的朋友、朋友的朋友，就能促成一次握手。 而网络交际，则大多数通过某些平台来实现，比如可以将自己放到一个平台中去，让很多 人看到，并且联系与认识。两者的优缺点很明显，社会交际的优点是可靠，彼此关系建立 在可靠的人际网络上，缺点是产生握手的时间长、代价较高。平台式的网络交际优点是成 本低，但不可靠。

**2.** **链接方式**

链接可以分为强链接和弱链接两类，能够充当信息桥的链接是弱链接，维系着群体、 组织内部的链接是强链接。弱链接在群体、组织之间建立了纽带联系。通过强链接获得的 信息往往重复性很高，而弱链接比强链接更能跨越其社会界限去获得信息和其他资源。

(1)弱链接

刚失业的A 在路上遇到了一年难得一见的熟人B, 两人聊起最近生活情况，A 对 B 说 自己正想找一个软件程序员的工作。B 突然想起了大学同学C 上周在一次聚会上提到他们 公司正在招聘，于是将C 的电话和电子邮件地址告诉了A。最后， A 通过C 应聘到了他们 公司。这个例子就体现了弱链接的威力，弱链接由六度分隔理论衍生而来。亲近的朋友之 间的关系是强链接，那么通过朋友的朋友，或者再隔几层关系就属于弱链接。六度分隔理 论说明了社会中普遍存在的弱链接关系，但是却可以发挥强大的作用，例如找工作时将体 会到这种弱链接的作用与效果。强弱关系还可根据建立关系的依据来决定，同兴趣、同群 组、同圈子、同应用，这类关系相对较弱，但同一类关系的交集越多，关系则可能会越强。

弱链接在与外界交流时发挥了关键的作用。为了得到新的信息，必须充分发挥弱链接 的作用。这些弱链接熟人是与外界沟通的桥梁，不同地方的人通过弱链接可以得到不同的 信息。而久不见面的人可能掌握了很多并不了解的情况。只有这些微弱关系的存在，信息 才能在不同的圈子中流传。每个人都有很多弱链接，这些弱链接在小世界中具有随意性和 极强的触发性，例如在自己的网站中添加朋友的链接或是关注的网站的链接是弱链接，然 后经常访问关注这些弱链接，这些弱链接转化为强链接。

(2)强链接

亲近的朋友之间常常互相交换信息，在连接循环中属于强链接，在这些关系循环中将 成员通过熟人的关系连接起来。而正是这些朋友的朋友，他们所提供的独特信息、机会可 能正是我们所需要的。弱链接在社会活动中扮演的角色不容忽视，尤其是在找工作时表现 得更加明显。

社交网络软件依据六度理论，以认识朋友的朋友为基础，扩展自己的人脉，并且无限 扩张自己的人脉，在需要的时候，可以随时获取一点，得到该人脉的帮助。社交网络的理 论基础是六度分隔理论。

**4.6.3** **150法则**

罗宾·邓巴指出大脑认知能力限制了特别物种个体社交网络的规模。根据猿猴的智力



**第** **4** **章** **大数据网络空间**

与社交网络推断，人类智力将允许人类拥有稳定社交网络的人数是148人。四舍五入大约 是150人，这就是著名的邓巴数字，又称之为150法则。

150法则指出人的大脑新皮层大小有限，提供的认知能力只能使一个人维持与大约150 个人的稳定人际关系，这一数字是人拥有的、与自己有私人关系的朋友数量。也就是说， 人们可能拥有1500名好友，甚至更多社交网站的好友，但只维持与现实生活中大约150个 人的内部圈子。而内部圈子好友在此理论中指一年至少联系一次的人。150 法则的含义是 每个人身后大致有150名亲朋好友。如果赢得了一个人的好感，就表明赢得了150个人的 好感；反之，如果得罪了一个人，也就表明了得罪了150 个人。在求职过程中，接触不同 的人，赢得对方的好感，求职者可以快速积累人脉资源，扩大人脉关系网。

150 法则在现实生活中的应用很广泛。比如，中国移动的“动感地带”SIM 卡只能保 存150个手机号， MSN 只能是一个 MSN 对应150个联系人。150成为普遍公认的可以与 之保持社交关系的人数的最大值。无论曾经认识多少人，或者通过一种社交网络服务与多 少人建立了弱链接，那些强链接仍然在此时此刻符合150法则。这也符合二八法则，即80% 的社会活动可能被150个强链接所占有。 一些研究表明，人们在多出这一数字的团体中合 作的效率会有所降低，人数太多不能进行有效的交流。这一法则也表明，当个体的生活圈 子过于狭小时就会感到孤独，个体需要他人的协助来发挥潜能。

**4.6.4** **唯象理论与唯象方法**

在这里，简单地介绍唯象理论与唯象方法。

**1.** **唯象理论**

唯象理论是知其然而不知其所以然的科学理论。唯象理论是实验现象更概括的总结和 提炼，但是无法用已有的科学理论体系做出解释与证明。“象”是中国古代哲学的重要范畴， 中国古代书画诗词强调意象，所以有“只可意会，不可言传”一说。唯象理论是说以“象” 为第一性，借助于现象或者直接从现象中来的理论。中医可以说是一种唯象理论，它已经 被几千年的生活实践所证明，但是却无法从物理、化学等现代科学角度进行解释。假说经 过证明就可以成为理论。

**2.** **唯象方法**

唯象方法是指在解释物理现象时，不追究微观原因，而是由经验总结和概括实验事实 得到自然界的基本规律做出演绎的推论。基于系统科学的观点，唯象方法是一种对简单巨 系统最常用的建模方法。它根据系统的宏观性质，不考虑系统的内部机制，直接利用系统 宏观层次上功能的特点建立演化方程。因此，用这种方法建立演化方程不必研究子系统之 间的相互作用。从方法论的角度看，唯象方法是通过客体宏观现象的实验观察，分析与归 纳找出规律，通过思维的创造，运用抽象与概括建立宏观的物理模型，利用数学工具构成 符合演绎逻辑的理论体系。





**大数据基础与应用**

**4.7** **社** **交** **网** **站**

**4.7.1** **社交网站作用**

社会网络是指个人之间的关系网络，基于社会网络关系系统思想的网站就是社交网站。 社会网络是指个人之间的关系网络，这种基于社会网络关系系统的网站就是社交网站。现 在许多社交网站都是 Web 2.0 网站，不仅一些大公司网站开始了一些社交网站应用， 一些 垂直领域的行业站点也开始了社交网站的应用，并获得了满意的效果。

在互联网中，由于 PC 机、智能手机都不具有强大的计算能力与带宽资源，所以只有 利用网站服务器，才能浏览发布信息。如果将每个设备的计算能力及带宽资源进行重新分 配与共享，这些设备就有可能获得比服务器更强大的能力，这就是分布计算理论诞生的根 源，是社交网站技术诞生的理论基础。

**4.7.2** **Web** **网站**

**1.Web** **2.0** **网站**

Web2.0 的概念始于2004年的O'Reilly 和 MediaLive International 之间的一场头脑风暴 论坛。头脑风暴是指一群人围绕一个特定的兴趣领域产生新观点的情境。由于在会议中使 用了没有拘束的规则，人们就能够更自由地思考，进入思想的新区域，从而产生很多的新 观点和问题解决方法。当参加者有了新观点和想法时，他们就大声说出来，然后在他人提 出的观点之上建立新观点。所有的观点被记录下来，但不进行批评。只有头脑风暴会议结 束的时候，才对这些观点和想法进行评估。头脑风暴的特点是让参会者敞开思想使各种设 想在相互碰撞中激起脑海的创造性风暴，其可分为直接头脑风暴和质疑头脑风暴法，前者 是在专家群体决策基础上尽可能激发创造性，产生尽可能多的设想的方法，后者则是对前 者提出的设想，方案逐一质疑，发行其现实可行性的方法，这是一种集体开发创造性思维 的方法。

Web 1.0网站的主要特点是用户仅通过浏览器获取信息，而Web 2.0 网站则更注重用户 的交互作用，用户既是网站内容的浏览者，也是网站内容的制造者。

**2.Web** **3.0** **网站**

Web 3.0的最大价值不是提供信息，而是提供基于不同需求的过滤器，每一种过滤器都 是基于一个需求。如果 Web2.0 解决了个性化的问题，那么 Web 3.0 就是解决最优化信息聚 合的问题。例如，用户只需要输入自己的需求，就可以迅速得到所需信息，甚至一套完整 的解决方案。

Web 3.0继承了 Web 2.0的以人为本理念，真正的Web 3.0 不仅止于根据用户需求提 供综合化服务，创建综合化服务平台，关键在于提供基于用户偏好的个性化聚合服务。





**第** **4** **章** **大数据网络空间**

在 Web 3.0 时代，同一模式化的综合门户将不复存在，例如新浪新闻首页是个人感兴趣的 新闻，而不感兴趣的新闻将不会显示。这种个性化的聚合必须依赖强大的智能化识别系统， 以及长期对于一个用户互联网行为规律的分析和锁定，它将颠覆传统的综合门户，使得 Web 3.0时代的互联网评价标准不再是流量和点击率，而是到达率和用户价值。因此， Web 3.0 是基于用户行为、习惯和信息的聚合而构建的，人性化、界面友好、简单易用一定是其核 心元素，基于用户需求的信息聚合才是互联网的趋势和未来。

小结

Web 2.0 网站的出现，为构建具有交互性的社交网站建立了基础。人们利用社交网站进 行一系列交互活动，促进了电子数据的迅猛发展，催生了大数据的出现。在互联网上大量 的、多类型的、非结构化的、变化快速的数据的产生、存储、传播与处理，需要全新的大 数据处理技术来完成。社会网络是大数据的一个重要研究方向，将计算科学与社会科学结 合，必将开拓一个崭新的研究领域，发展前景远大。



Hadoop发展历程

分布式系统与Hadoop

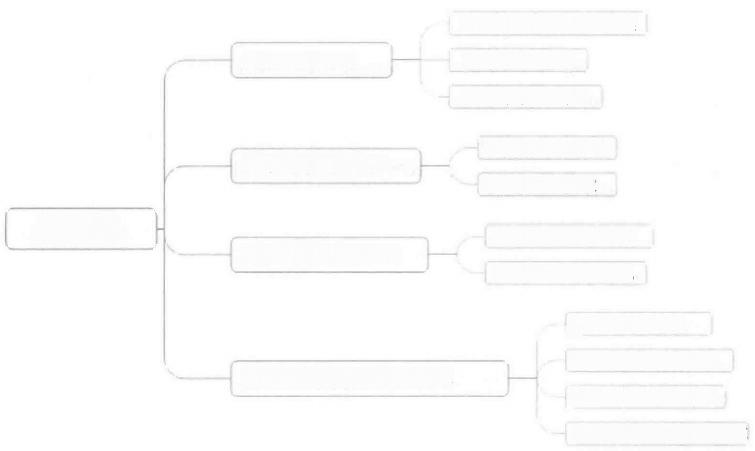
基于Hadoop平台的分布式计算

SQL 数据库和Hadoop

基于Hadoop的分布式计算

**第5章** **MapReduce 分布编程模型**

本章主要内容



函数型语言与函数式编程

函数式编程优点

函数式编程的特征

映射与映射函数

映射函数与化简函数

化简与化简函数

MapReduce计算描述

MapReduce的体系结构

MapReduce适用情况

MapReduce 分布编程模型

函数式编程范式

计算机科学是算法与算法变换的科学，算法是计算机科学的基石。任何一个计算问题 的分析与建模，几乎都可以归为算法问题。MapReduce 算法模型是由 Google 公司针对大 规模群组中的海量数据处理而提出的分布编程模型，主要应用于大规模数据集(大于1TB) 的分布并行运算。在 MapReduce 模型中的 Map (映射)和 Reduce (化简)创意来自函数 型编程语言，同时也继承了向量型编程语言的特性。MapReduce 模型能够使程序员在不了 解分布式并行编程的情况下，将自己书写的程序在分布式系统上运行。MapReduce 模型主 要实现了映射与化简两个核心功能，分别由Map 和 Reduce 完成，这两个函数也是函数型 语言中常用的函数，由用户负责实现。Map 函数应用于集合中的所有成员，然后返回一个 基于这个处理的结果集。而 Reduce 函数是从两个或更多个 Map 结果中，通过多个线程、 进程或者独立系统并行执行处理的结果集进行分类和归纳。 一个 Map 函数用来把一组键值 对映象成一组新的键值对，Reduce 函数用来对同一个键的值进行合并。

**5.1** **函数式编程范式**

函数式编程是一种编程范式，在这种编程范式中，更多地使用了函数运算。



第 5 章 MapReduce 分布编程模型

**5.1.1** **函数型语言与函数式编程**

函数型语言是一种典型的程序设计语言。函数型语言的特点是把问题求解过程表示成 块结构，对调用块的调用者来说，每个块都有输入数据和经过加工处理后的输出数据。每 个块的功能就如同函数的功能，例如 LISP 语言、ML 语言等都属于函数型语言。

利用函数型语言编程就是一种函数式编程，函数式编程使用一系列的函数解决问题。 任何一个函数仅接受输入数据并产生输出函数值，不包含任何能影响产生输出的内部状态。 任何情况下，使用相同的参数调用函数始终能得到一致的结果。

在一个函数式的程序中，输入数据要通过一系列的函数，每个函数根据它的输入产生 输出。函数式编程位于面向对象编程的对立面。面向对象编程通常包含内部状态(字段) 和许多能修改这些状态的函数，程序运行中则不断修改状态。函数式编程则极力避免状态 改动。对于复杂的系统，可以同时使用函数式编程和面向对象编程，在这种混合应用中， 既采用了面向对象技术建模，还获得了函数式编程风格的优点。

**5.1.2** **函数式编程优点**

函数式编程具有如下优点。

**1.** **逻辑可证**

由于没有边界效应，所以更容易从逻辑上证明程序正确性，而不是仅通过测试完成。

**2.** **模块化**

函数式编程以简单为原则， 一个函数只做一件事情，将大的功能模块拆分成尽可能小 的功能模块。小的功能模块更易于阅读和检查错误。

**3.** **组件化**

更容易利用简单的小函数加以组合形成新的功能更强的函数，体现了组件化的特点。

**4.** **易于调试**

细化的、定义清晰的函数能够使得调试更加简单。当程序运行不正常时，每个函数 都是检查数据是否正确的连接口，能更快速地排除没有问题的代码，定位到出现问题的 地方。

**5.** **易于测试**

函数不依赖于系统状态，无须在测试前构造测试桩，更加容易编写单元测试的代码。

**6.** **更高的生产率**

函数式编程产生的代码比其他技术更简洁，往往是其他技术的一半左右，并且更容易 阅读和维护。



**大数据基础与应用**

**5.1.3** **函数式编程的特征**

函数式编程具有下述特征。

**1.** **没有副作用**

函数程序由一系列函数对数据的变换构成，因为函数式编程中的每个符号都是最终的， 所以从来没有在某个地方修改过值，也没有函数修改过在其作用域之外的量并被其他函数 使用。这表明函数求值的结果只是其返回值，而唯一影响其返回值的就是函数的参数，函 数操作数据，所以函数没有产生副作用。

**2.** **无状态的编程**

在现实情况中，状态不可能一直保持不变，而状态必然需要改变与传递，那么在函数 式编程中的则是将其保存在函数的参数中，作为函数的附属品来传递。

**3.** **输入值和输出值**

在函数式编程中，只有输入值和输出值。函数是基本单位，几乎被用作所有部分，包 括最简单的计算，甚至连变量都被计算所取代。变量只是一个名称，而不是一个存储单元， 这是函数式编程与传统的命令式编程最典型的不同之处。

在面向对象编程中，将对象传来传去。在函数式编程中，是将函数传来传去，又将这个 函数称为高阶函数。在数学和计算机科学中，高阶函数是至少满足下列一个条件的函数。

① 接受一个或多个函数作为输入；

② 输出一个函数值。

**5.2** **映射函数与化简函数**

在程序语言中，映射函数与化简函数是常用的函数。映射函数与化简函数是 MapReduce 分布编程模型的两个主要函数。简单地说， 一个映射函数就是对一些独立元素组成的概念 上的列表(例如， 一个测验成绩的列表)的每个元素进行所指定的操作。事实上，每个元 素都是被独立操作的，而原始列表没有被更改，因为这里创建了一个新的列表来保存新的 答案。这就可以看出，Map 操作可以高度并行，非常适用分布并行计算领域。而化简操作 指的是对一个列表的元素进行适当的合并，虽然不如映射函数那么高度并行，但是因为化简 总是有一个简单的答案，大规模的运算相对独立，所以化简函数也适用于高度并行环境。

**5.2.1** **映射与映射函数**

**1.** **映射**

设A 、B 是两个非空集合，如果存在一个法则f, 使得对A 中的每个元素a, 按法则f,





第 5 章 MapReduce 分布编程模型

在B 中有唯一确定的元素b 与之对应，则称f 为从A 到 B 的映射，记作f:A→B。其中， b 称为元素a 在映射f 下的象，记作b=f(a),a 称为b 关于映射f 的原象。集合A 中所有元 素的象的集合成为映射f 的值域，记作f(A)。

映射还可以用于定义函数。函数是从非空数集到非空数集的映射，而且只能是一对一 映射或多对一映射。按照映射的定义，下面的对应是映射。

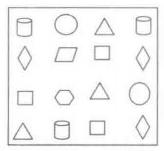
设A={1,2,3,4},B={3,5,7,9}, 集合A 中的元素x 按照对应关系“乘2加1”和集合B 中的元素对应，这个对应是集合A 到集合B 的映射。

映射在不同的领域有很多的名称，例如函数、算子等。

**2.** **映射函数**

映射函数的功能描述如图5- 1 所示。左排表示原始输入列表，即输入的是形状图形； 右排表示输出列表，即由两个处理单元完成各种图形的数量统计；中间箭头表示映射函数， 即 Map 函数就是一个图形统计函数。得到的结构如下：上面的处理单元1的统计结果是圆 形1个、菱形1个、正方形3个、圆柱形2个、六角形1个。下面的处理单元2的统计结 果是圆形1个、菱形2个、正方形1个、三角形2个、圆柱形1个、平行四边形1个。事 实上，每个图形被独立操作，但原始列表没有被更改，而且创建了一个新的列表来保存新 的答案。这表明，映射操作可以高度并行，非常适用于高性能要求的应用以及分布、并行 计算领域的问题需求。

输入数据集



处理单元1输出结果

Map

处理单元2输出结果

|  |
| --- |
| 2 |
|  |
| 2 |

图5-1 映射函数的功能举例

**5.2.2** **化简与化简函数**

**1.** **化简**

化简在数学上是一个非常重要的概念。化简一般指把复杂式子化为简单式子的过程， 在数学运算中，必须通过化简才能简便地求出值。分式化简称为约分。整式化简包括移项、





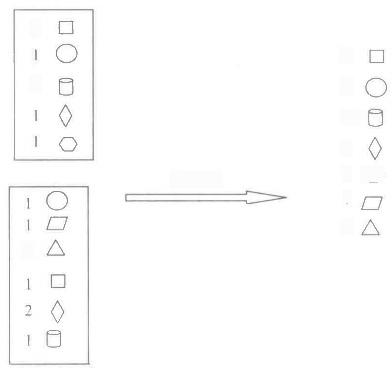
**大数据基础与应用**

合并同类项、去括号等，化简后的式子一般为更简式，项数减少。

例如，解方程也可以看作是一个化简的过程。

**2.** **化简函数**

化简操作能够完成对一个列表的元素进行适当的合并。例如，如果将Map 函数的输出 作为 Reduce 函数的输入，可以使用一个处理单元完成化简工作，那么 Reduce 函数的输出 结果如图5-2所示。其中正方形4个、圆形2个、圆柱形3个、菱形3个、六角形1个、 平行四边形1个、三角形2个。这是对输入的16个图形的最后统计结果，先进行映射函数 作用，后进行化简函数作用。化简总是有一个简单的答案，化简操作是大规模的相对独立 运算，所以化简函数也适合于高度并行环境中使用。

3

4

2

2

3

[

3

1

〇

Reduce

1

2

2

图5-2 化简函数的功能举例

**5.3** **MapReduce** **的体系结构**

MapReduce 是针对大规模数据集并行计算而设计的编程模型，在 Hadoop 分布计算平 台中，利用 MapReduce 模型对任务进行分配，进而使分配后的任务在大量计算机组成的集 群上进行并行计算，实现了Hadoop 对任务的并行处理的功能。

**5.3.1** **MapReduce** **计算描述**

MapReduce 以函数方式提供了Map 和 Reduce 来进行分布式计算。Map 相对独立且并 行运行，对存储系统中的文件按行处理，并产生键值对 (key/value) 。Reduce 以 Map 的输 出作为输入，相同 key的记录汇聚到同一个 Reduce,Reduce 对这组中间结果进行操作，将 中间结果相同的键进行合并规约，并产生最终结果，即产生新的数据集。所有 Reduce 任务 的输出组成最终结果。形式化描述如下。



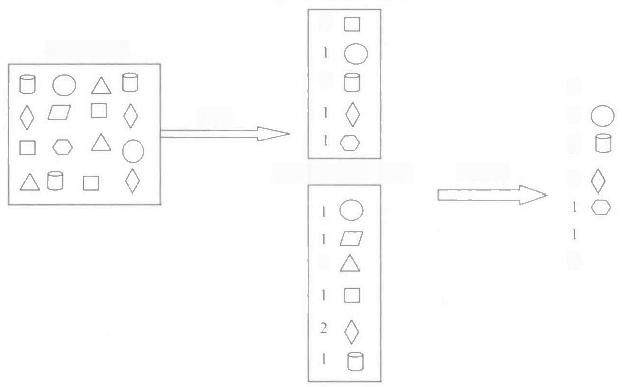
第 5 章 MapReduce 分布编程模型

Map:(k1,v1)->list(k2,v2)

Reduce:(k2,list(v2))>list(v3)

如图5-2所示的是以图5-1 所示的右排输出作为输入，右排为化简函数的输出。不难 看出，经过 Map 与 Reduce 两步计算，可以完成形状计数。而基于 MapReduce 分布计算模 型的形状计数全过程如图5-3所示。

处理单元1输出结果

3

输入数据集

2

4

2

Map

3

处理单元2输出结果

Reduce

3

2 2

图5-3 基于MapReduce 的形状计数

映射和化简函数的功能是按一定的映象规则将输入的键值对转换成另一个或另一批键 值对输出，如表5-1 所示。

**表5-1** **映射函数和化简函数功能说明**

|  |  |  |  |
| --- | --- | --- | --- |
| **函** **数** | **输** **入** | **输** **出** | **说** **明** |
| 映射 | <k1,vl> | List(<k2,v2>) | (1)将数据集解析成一批<key,value>对，输入Map 函数中进行处理  (2)对于每一个输入的<k1,v1>,将输出一批<k2,v2>, <k2,v2>是计算的中间结果 |
| 化简 | <k2,List(v2)> | <k3,v3> | 输入的中间结果<k2,List(v2)>中的List(v2)表 示属于同一个k2的value |

Reduce 函数的作用是把大的数值列表转变为一个(或几个)输出数值。在 MapReduce 中，所有的输出数值一般不会被 Reduce 集中在一起。有着相同键的所有数值将被一起送到 一个 Reduce 中。作用在有着不同键关联的数值列表上的 Reduce 操作之间独立执行，如图5-4 所示。

Reduce 一般用来生成总结数据，把大规模的数据转变成更小的数据信息。例如“+” 可以用来作为一个 Reducing 函数，用于返回输入数据列表的值的总和。基于 MapReduce



**大数据基础与应用**

计算模型的分布式并行程序设计非常简单，程序员的主要工作就是编写Map 函数和 Reduce 函数的代码。其他的并行编程中的种种复杂问题，如分布式存储、工作调度、负载平衡、 容错处理、网络通信等，均由Hadoop 平台负责处理。也就是说，这些内容对程序员透明。

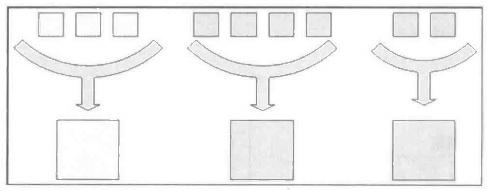


图5-4 多个Reduce 任务

**5.3.2** **MapReduce 适用情况**

由于MapReduce 编程模型是对输入按行依次处理，所以它更适用于对批量数据进行处 理。由于良好的横向可扩展性， MapReduce 尤其适用于对大规模数据的处理。但是，对搜 索等只是需要从大量数据中选取某几条进行特别的操作，相对于具有完善索引的系统而言， MapReduce 不具有优势。因为它需要对每条数据进行匹配，并把与搜索条件相匹配的数据 提取出来。而如果采用索引系统，并不需要遍历所有的数据。另外，由于每次操作需要遍 历所有数据，MapReduce 并不适用于需要实时响应的系统。相反地，对于搜索引擎的预处 理工作，比如网页爬虫、数据清洗以及日志分析等实时性要求不高的后台处理工作， MapReduce 编程模型足以胜任。

**5.4** **基于** **Hadoop** **平台的分布式计算**

分布式计算是一个广义的概念，在计算机科学中，有多种方式支持分布式计算。

**5.4.1** **Hadoop** **发展历程**

Hadoop 技术是推动大数据应用的引擎，用于收集、共享和分析来自网络的大量结构化、 半结构化和非结构化数据。应用 Hadoop 技术之前需要做一些技术准备。利用一定的时间 确定需要处理的数据路线图，认真研究 Hadoop 技术如何与网络的其他部分相配合，开发 一个明确的分类学模型。Hadoop 开源框架的发展历程的描述如图5-5所示。

**1.Hadoop 框架**

Hadoop 是一个开源框架，它实现了 MapReduce 分布算法，用以查询在互联网上的分 布数据。在 MapReduce 算法中， Map 的功能是将查询操作和数据集分解成组件， Reduce 的功能是在查询中映射的组件可以被同时处理(即归约),从而可以快速地返回结果。



照称

2001年著,互联阿迅

速发展。数细量成作递 增，量变引起质安，开

世理技木提出

全新的要求

数据位率

共苏型

数据仓率开

始两现，关

加

百被发

且平台 关，

课人学数 期

第一

模型

速发1

世

1946 1951 1956 1961 1970 19741979

★

20世纪60年代，Ⅱ系统规植和复杂变变大，数据与应用分 离的号求开始产生，数招库技术开始萌芽并蓝的发展，

并在1990午后还步统一到以关系星数招库为主导

1946年，电脑诞生，数 拓与应用案密烟纳在文件 中，独此不分

Hadooo 或 方APocha 道级项日。

支 持

台称发表

市式计

SQ 语

E发布第

个50高

关1系数

后读

算

三数据

重

雨

被 重

管理

dog

月第一个对 心

但汉； 的主切 2

理

1991 20012003 2008 2011



第5章 MapReduce 分布编程模型

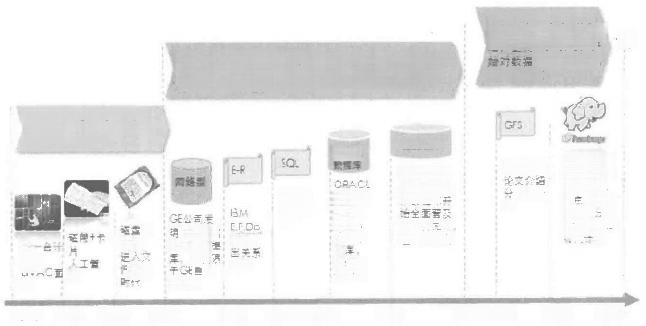


图5-5 Hadoop 开源框架的出现历程

**2.Hadoop 的** **主** **要** **特** **点**

(1)方便

Hadoop 运行在由多机构成的大型集群上，或者云计算平台等云计算服务上。适于运行 大型分布式程序。

(2)健壮

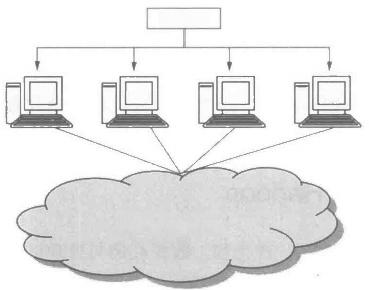
如果架构硬件频繁地出现故障，那么 Hadoop 可以处理大多数此类故障，进而可以胜 任更严苛的工作。

(3)横向可扩展

Hadoop 通过增加集群节点，可以线性地横向扩展以处理更大的数据集。

(4)简单

Hadoop 允许用户快速编写出高效的并行代码，进而可以廉价地建立 Hadoop 集群。图5-6 所示的是用户与Hadoop集群的交互解释说明。 一个 Hadoop 集群拥有许多并行的计算机， 用以存储与处理大数据。客户端计算机发送作业到集群云并获得结果，实现了以计算为中 心到以数据为中心的转变。

客户端

Hadoop集群云

图5- 6 用 户 与 Hadoop 集群云的交互解释说明





**大数据基础与应用**

Hadoop 集群云是指在同一地点用网络互联的一组通用机器。数据存储与处理都发生在 这些机器云中，不同的用户可以从独立的客户端提交计算作业到 Hadoop 集群云，这些客 户端可以是远离 Hadoop 集群云的个人计算机。但是分布式系统的构建存在多种形式，并 不仅如图5-6所示。

**5.4.2** **分布式系统与** **Hadoop**

在解决大规模计算问题时，不能单纯依赖于制造越来越大型的服务器(即纵向扩展)。 横向扩展显示出了强大的灵活性，并已经获得广泛应用，即把许多低端的机器组织在一起， 形成一个功能专一的分布式系统。

横向扩展的分布式系统与纵向扩展大型单机服务器之间的比较，需要考虑现有 I/O 技 术的性价比。对于一个有4个I/O 通道的高端机，即使每个通道的吞吐量各为100 MB/s, 读取4TB 的数据集也需要大约3个小时。而利用 Hadoop, 同样的数据集会被划分为较小 的块(通常为64MB), 通过Hadoop 分布式文件系统 (HDFS) 分布在集群内多台机器上。 使用适度的复制，集群可以并行读取数据，进而提供很高的吞吐量。而这样一组通用机器 比一台高端服务器更加便宜。

前面的解释充分展示了 Hadoop 相对于单机系统的高效率。现在将 Hadoop 与其他分布 式系统架构进行比较。 一个众所周知的例子是利用普适计算来协助寻找外星生命。利用一 台中央服务器存储来自太空的无线电信号，并在网上发布给世界各地的客户端台式机去寻 找异常的迹象。这种方法是以计算为中心的方法，将数据移动到计算即将发生的地方，经 过计算后，再将返回的数据结果存储起来。

Hadoop 与普适计算不同。普适计算需要客户端和服务器之间重复地传输数据。这虽然 能够很好地适应数据密集型的计算工作，但由于数据规模太大，数据移动变得十分困难。 Hadoop 强调将代码向数据迁移，而不是相反。从图5-6 中可以看出，Hadoop 的集群内部 既包含数据又包含计算环境。客户端仅需发送待执行的MapReduce 程序，而这些程序代码 一般都很小(通常为几千字节)。更重要的是，处理程序向数据迁移的理念被应用在Hadoop 集群自身。数据被拆分后分布于集群中，并且尽可能使对一段数据的计算发生在同一台机 器上，即这段数据驻留的地方。

程序代码向数据迁移的理念符合 Hadoop 面向数据密集型处理的设计目标。要运行的 程序规模比数据小几个数量级，更容易移动。此外，在网络上移动数据要比在其上加载代 码花费更多时间。因此，数据不移动，而将可执行代码移动到数据所在的机器上去，这就 是常说的以数据为中心的理念。

**5.4.3** **SQL** **数据库和** **Hadoop**

Hadoop 是一个数据处理框架。在当前，数据存储与处理的主要工具是标准的关系数据 库。Hadoop 的优势是：SQL (结构化查询语言)是针对结构化数据而设计，而 Hadoop 最 初的许多应用是针对非结构化数据，显然，Hadoop 比 SQL 的应用更为广泛。





第 5 章 MapReduce 分布编程模型

如果只对结构化数据处理，则Hadoop和SQL 需要做更深入的比较。原则上，SQL 和 Hadoop 可以互补，因为 SQL 是一种查询语言，可将 Hadoop 作为其执行引擎。但实际上， SQL 数据 库往往指一整套传统技术，许多这类关系数据库无法满足 Hadoop 设计所面向的需求。

考虑到这一点，Hadoop 与 SQL 数据库的比较如下。

**1.** **横向扩展代替纵向扩展**

关系型数据库更容易纵向扩展。要运行一个更大的数据库，就需要买一个更大的机器。 事实上，往往会看到服务器厂商在市场上将其昂贵的高端机标称为数据库级的服务器。不 过有时可能需要处理更大的数据集，却找不到一个足够大的机器。更重要的是，高端的机 器对于许多应用并不经济。例如，性能4倍于标准 PC 的机器，其成本将大大超过放在一 个集群中的标准4台PC 。Hadoop 的设计就是为了能够在PC 集群上实现横向扩展的架构。 添加更多的资源，对于Hadoop 集群就是增加更多的机器。 一个 Hadoop 集群的标配是十至 数百台计算机。事实上，如果不是为了产品开发，没有理由在单个服务器上运行 Hadoop。

**2.** **键值对代替关系表**

关系数据库的一个基本原则是让数据按某种模式存放在具有关系型数据结构的表中。 虽然关系模型具有大量形式化的属性，但是许多当前的应用所处理的数据类型并不能很好地 适应这个模型。文本、图片和 XML 文件就是最典型的例子。此外，大型数据集往往是非结 构化或半结构化的。Hadoop 使用键值对作为基本数据单元，可足够灵活地处理较少结构化 的数据类型。在 Hadoop 中，数据的来源可以有任何形式，但最终会转化为键值对处理。

**3.** **函数式编程代替声明式查询**

SQL 是一种高级声明式语言，查询数据的手段是声明需要的查询结果并使数据库引擎 判定如何获取数据。在 MapReduce 编程模型中，数据处理步骤由用户指定，类似于 SQL 引擎的一个执行计划。SQL 使用查询语句，而 MapReduce 则使用脚本和代码。MapReduce 可以用比 SQL 查询更为一般化的数据处理方式。例如，可以建立复杂的数据统计模型或改 变图像数据的格式，而 SQL 就不能很好地适应这些任务。

另一方面，当数据处理非常适合于关系型数据结构时，可以发现使用 MapReduce 并不 自然。习惯于 SQL 范式的人使用 MapReduce 是一个新的挑战。更轻松地掌握 MapReduce 编程并非易事，这里还有很多扩展可用，便于人们采用更熟悉的范式来编程，同时拥有 Hadoop 的可扩展性优势。事实上，使用某些扩展可采用一种类似 SQL 的查询语言，并在 MapReduce 上运行。

**4.** **离线批量处理代替在线处理**

Hadoop 是专为离线处理大规模数据分析而设计，它并不适合对几个记录随机读写的在 线事务处理模式。事实上，Hadoop 最适合一次写入、多次读取的数据存储需求。在这方面 它与数据仓库相同。

Hadoop 是一个通用的工具，它使新用户可以享受到分布式计算的好处。通过采用分布 式存储、迁移代码而非迁移数据，Hadoop 在处理大数据集时避免了耗时的数据传输问题。



**大数据基础与应用**

此外，数据冗余机制允许 Hadoop 从单点失效中恢复。在 MapReduce 框架中编写程序非常 方便，由Hadoop分配任务到执行节点、管理节点间的通信，用户不必操心如何分割数据等。

**5.4.4** **基于** **Hadoop的分布式计算**

基于 MapReduce 框架所设计的计算是分布式计算。在 Hadoop 中，分布式文件系统为 各种分布式计算需求服务。分布式文件系统就是增加了分布式的文件系统，将定义推广到 类似的分布式计算上，可以将其视为增加了分布式支持的计算函数。从计算的角度上看， MapReduce 框架接受各种格式的键值对文件作为输入，读取并计算后，最终生成自定义格 式的输出文件。而从分布式的角度来看，分布式计算的输入文件往往规模巨大，并且分布 在多个机器上，单机计算完全不支撑并且效率低下，因此 MapReduce 框架需要提供一套机 制，将计算扩展到无限规模的机器集群上进行。

在 MapReduce 框架中，将每一次计算请求称为作业。作业可分为两步骤完成。首先是 将其拆分成若干个 Map 任务，分配到不同的机器上去执行，每一个 Map 任务将输入文件 的一部分作为自己的输入，经过一些计算，生成某种格式的中间文件，这种格式与最终所 需的文件格式完全一致，但是仅仅包含一部分数据。因此，等到所有 Map 任务完成后，进 入下一个步骤，用以合并这些中间文件获得最后的输出文件。此时，系统会生成若干个 Reduce 任务，同样也是分配到不同的机器去执行，其目标就是将若干个Map 任务生成的中 间文件汇总到最后的输出文件中，这就是 Reduce 任务的价值所在。经过如上步骤后，作业 完成，所需要的目标文件生成。算法的关键就在于增加了一个中间文件生成的流程，大大 提高了灵活性，使其分布式扩展性得到了保证。

Hadoop 术语的解释如表5-2所示。

**表5-2** **术语解释**

|  |  |  |
| --- | --- | --- |
| **Hadoop中文术语** | **Hadoop英文术语** | **相** **关** **解** **释** |
| 作业 | Job | 用户的每一个计算请求，就称为一个作业 |
| 作业服务器 | Master | 用户提交作业的服务器，同时，它还负责各个作业任务的分 配，管理所有的任务服务器 |
| 任务服务器 | Worker | 负责执行具体的任务 |
| 任务 | Task | 每一个作业，都需要拆分开，交由多个服务器来完成，拆分 出来的执行单位，就称为任务 |
| 备份任务 | Buckup Task | 每一个任务，都有可能执行失败。为了降低为此付出的代价， 系统会在另外的任务服务器上执行同样一个任务，这就是备 份任务 |

**1.** **基本架构**

在 Hadoop 架构中，作业服务器称为 Master。作业服务器负责管理运行在此框架下的

一个作业的计算流程如图5-7所示。

在每个任务的执行中，又包含输入的准备、算法的执行、输出 的生成三个子步骤。按照这个流程，可以很清晰地了解整个 MapReduce 框架下作业的执行过程。

(1)作业提交

一个作业在提交之前需要完成所有配置，因为一旦提交到了作 业服务器上，就进入了完全自动化的流程。用户除了观望，最多也 就能起一个监督作用。用户在提交代码阶段，需要做的主要工作是 写好所有自定的代码，主要有Map 和 Reduce 的执行代码。

(2)Map 任务的分配

当一个作业提交到了作业服务器上，作业服务器将生成若干个 Map 任务，每一个 Map 任务负责将一部分的输入转换成格式与最终 格式相同的中间文件。通常一个作业的输入都是基于分布式文件系 统的文件，而对于一个 Map 任务而言，它的输入往往是输入文件的



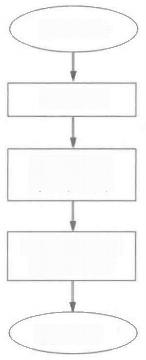
**第** **5** **章** **MapReduce** **分布编程模型**

所有作业，也是为各个作业分配任务的核心。与 HDFS 的主控服务器类似，简化了负责的 同步流程。执行用户定义操作的是任务服务器，每一个作业被拆分成多个任务，包括 Map 任务和 Reduce任务等。任务是执行的基本单元，它们都需要分配到合适任务服务器上去执 行，任务服务器一边执行一边向作业服务器汇报各个任务的状态，以此来帮助作业服务器 了解作业执行的整体情况，以及分配新的任务等。

除了作业的管理者与执行者之外，还需要一个任务的提交者，这就是客户端。与分布 式文件系统一样，用户需要自定义好所需要的内容，经由客户端相关的代码，将作业及其 相关内容和配置提交到作业服务器，并随时监控执行的状况。

与分布式文件系统相比，MapReduce 框架还有一个特点就是可定制性强。文件系统中 很多的算法都很固定和直观，不会由于所存储的内容不同而有太多的变化。而作为通用的 计算框架，需要面对的问题则更复杂。在不同的问题、不同的输入、不同的需求之间，很 难存在一种通用的机制。对于MapReduce 框架而言， 一方面要尽可能抽取出公共需求，并 实现它；另一方面要提供良好的可扩展机制，满足用户自定义各种算法的需求。

**2.** **计算流程**



作业开始

作业提交

Map任务的

分配和执行

Reduce任务的

分配和执行

作业完成

图5-7 作业执行过程

一个数据块，或者是数据块的一部分，但通常不跨数据块。因为一旦跨了数据块，就可能 涉及多个服务器，带来不必要的麻烦。

当一个作业从客户端提交到了作业服务器上，作业服务器将作业拆分成若干个Map 任 务后，会预先挂在作业服务器上的任务服务器拓扑树上。这是依照分布式文件数据块的位 置来划分的，比如一个Map 任务需要用某个数据块，这个数据块有三份备份，那么在这三 台服务器上都会挂上此任务，可以视为一个预分配。

任务分配是一个重要的环节，任务分配就是将合适的作业分配到合适的服务器上。主 要分为两个步骤：

① 选择作业，然后是在此作业中选择任务。与所有分配工作一样，任务分配也是一个



**大数据基础与应用**

复杂的工作。不当的任务分配，可能会导致网络流量增加、某些任务服务器负载过重、效 率下降等。不仅如此，任务分配还无一致模式，不同的业务背景，可能需要不同的分配算 法才能满足需求。当一个任务服务器工作得游刃有余，期待获得新的任务的时候，将按照 各个作业的优先级，从最高优先级的作业开始分配。每分配一个，还会为其留出余量，以 备不时之需。举一个例子：系统目前有优先级3、2、1的三个作业，每个作业都有一个可 分配的Map 任务， 一个任务服务器来申请新的任务，它还有能力承载3个任务的执行，将 先从优先级为3的作业上取一个任务分配给它，然后再留出一个任务的余量。此时，系统 只能再将优先级为2作业的任务分配给此服务器，而不能分配优先级1的任务。这样的策 略，基本思路就是一切为优先级高的作业服务。

② 确定了从哪个作业提取任务后，具体的分配算法很简单，就是尽全力为此服务器分 配任务，也就是说，只要还有可分配的任务，就一定会分给它，而不考虑后来的服务器。 作业服务器会从离它最近的服务器开始，看上面是否还挂着未分配的任务(预分配上的), 从近到远，如果所有的任务都分配了，那么看有没有开启多次执行，如果开启，考虑把未 完成的任务再分配一次。

对于作业服务器来说，把一个任务分配出去了，并不意味着作业服务器工作完成，对 此任务可以不管不顾了。因为任务可以在任务服务器上执行失败，可能执行缓慢，这都需 要作业服务器帮助它们再来执行一次。

(3)Map 任务的执行

与 HDFS 类似，任务服务器是通过心跳消息向作业服务器汇报此时各个任务执行的状 况，并向作业服务器申请新的任务。在实现过程中，使用池化的方式。有若干个固定的槽， 如果槽没有满，那么就启动新的子进程，否则，就寻找空闲的进程。如果是同任务的直接 放进去，否则杀死这个进程，用一个新的进程代替。每一个进程都位于单独线程中。但是 从实现上看，这个机制好像没有部署，子进程是死循环等待，而不会阻塞在父进程的相关 线程上。父线程的变量一直都没有调整， 一旦分配，始终都处在繁忙的状况。

(4)Re duce 任务的分配与执行

Reduce 的分配比 Map 任务简单，基本上是所有 Map 任务完成了，如果有空闲的任务 服务器，就给分配一个任务。因为Map 任务的结果星罗棋布，且变化多端，真要搞一个全 局优化的算法，得不偿失。而Reduce 任务执行进程的构造和分配流程，与Map 基本一致。 但 Reduce 任务与Map 任务的最大不同是 Map 任务的文件都存于本地，而 Reduce 任务需 要到处采集。这个流程是作业服务器经由此 Reduce 任务所处的任务服务器，告诉 Reduce 任务正在执行的进程，它需要的Map 任务执行过的服务器地址，此 Reduce 任务服务器会 与原Map 任务服务器联系，通过FTP 服务下载。这个隐含的直接数据联系，就是执行Reduce 任务与执行Map 任务最大的不同了。

(5)作业的完成

当所有 Reduce 任务都完成之后，所需数据都写到了分布式文件系统上，整个作业才正 式完成了。

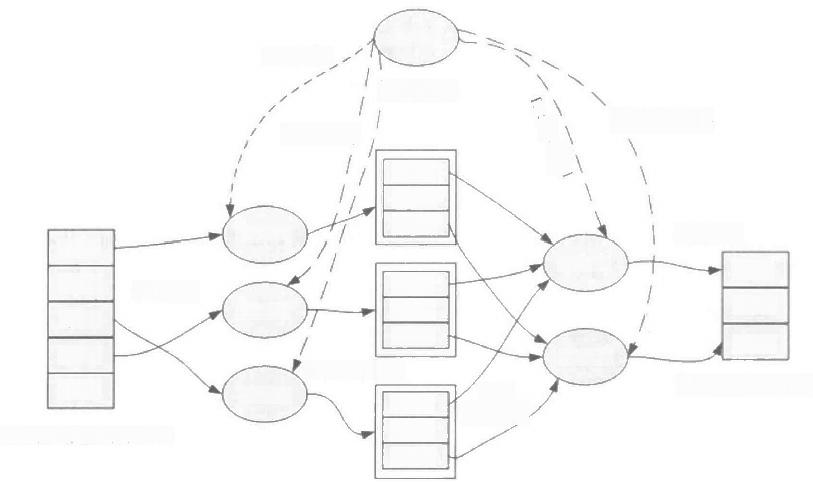
**3.MapReduce 程序的执行过程**

基于MapReduce 算法编写的 MapReduce 程序的分布式执行过程如图5-8所示。





第 5 章 MapReduce 分布编程模型

分片1 分片2 分片3 分片4

分片5

论文集，分成M份

读取数据

,M=5

Master

Map任务1

|Map 任务3

Map任务4/

分区1 分区2 分区3

Worker

分区1 分区2 分区3

Worker

写入本地磁盘

分区1 分区2 分区3

Worker

中间键值对

|  |  |
| --- | --- |
| 22  管 | Reduce 任务3 |

|  |  |
| --- | --- |
| Worker  Worker  读取  中间数据 | 输出结果  输出1  输出2  输出3  输出分为R份 ，R=3 |

图5 - 8 MapReduce 程序的执行过程

用户程序中的MapReduce 类库首先将输入文档分割成大小为16MB～64MB 的文件片 段，用户也可以通过设置参数对大小进行控制。随后，集群中的多个服务器开始执行多个 用户程序的副本。

由 Master 负责分配任务，如果总计分配M 个 Map 任务和R 个 Reduce 任务。分配的 原则是 Master 选择空闲的 Worker并为其分配一个 Map 任务或一个 Reduce 任务。

被分配到Map 任务的 Worker 读取对应文件片段，从输入数据中解析出键值对，并将 其传递给用户定义的Map 函数。由Map 函数产生的键值对被存储在内存中。

缓存的键值对被周期性写入本地磁盘，并被分成R 个区域。这些缓存数据在本地磁盘 上的地址被传递回Master, 由 Master 再将这些地址送到负责Reduce 任务的 Master。

当负责 Reduce 任务的 Master 得到关于上述地址的通知时，它使用远程过程调用从本 地磁盘读取缓冲数据。随后Worker 将所有读取的数据按键排序，使得具有相同键的对排在 一起。

对于每一个唯一的键，负责 Reduce 任务的 Worker 将对应的数据集传递给用户定义的 Reduce 函数。这个Reduce 函数的输出被作为 Reduce 分区的结果添加到最终的输出档中。

当所有的 Map 任务和 Reduce 任务都完成时，Master 唤醒用户程序。此时，用户程序 的 MapReduce 调用向用户的代码返回结果。

MapReduce 模型通过将数据集的大规模操作分发给网络上的各节点实现可靠性，每个 节点将完成的工作和状态更新周期性地报告。如果一个节点保持沉默超过一个预设的时限， 主节点记录下这个节点状态为死亡，并把分配给这个节点的数据发到别的节点。每个操作 使用原子操作以确保不会发生并行线程间的冲突 ， 当文件被改名的时候 ， 为了避免副作用 ，



**大数据基础与应用**

系统将它们复制到任务名以外的另一个名字上去。

由于化简操作并行能力较差，主节点会尽量把化简操作调度在一个节点上，或者离需 要操作的数据尽可能近的节点上。这种做法适用于具有足够的带宽、内部网络没有那么多 的机器情况下的需求。

**4.** **举例**

(1)键和值

在 MapReduce 中，每一个值都有一个键与其关联，键标识相关的值。例如，从多辆车 中读取到的时间、车速表日志可以由车牌号码(键)标识的内容(车速和时间)如下：



AAA-12365mph,12:00pm

ZZZ-78950mph,12:02pm

AAA-12340mph,12:05pm

CCC-45625mph,12:15pm

mapping 和 reducing 函数接收键值对。这些函数的每一个输出都是一样的，都是一个 键和一个值，并将被送到数据流程的下一个列表。mapp 针对每一个输入元素都要生成一个 输出元素，reduce 针对每一个输入列表都要生成一个输出元素。但在 MapReduce 中，每一 个阶段都可以生成任意的数值， mappr 可能把一个输入map 生成0个、1个或100个输出， reduce 可能计算超过一个输入列表并生成一个或多个不同的输出。

(2)词频统计

统计一组文档中的每个单词出现的次数，在这个例子中的文档有一个文件，每个文件 中只有一句话。

Do as I say,not as l do.

将得到如表5-3所示的单词统计值。

**表5-3** **单词统计值**

|  |  |  |  |
| --- | --- | --- | --- |
| 单 词 | **计** **数** | 单 词 | 计 数 |
| as | 2 | not | 1 |
| do | 2 | say | 1 |
| i | 2 |  |  |

将这个特定的练习称为单词统计。如果文档集合很小， 一个简单的程序即可完成这项 工作。可写为如下一段伪代码：



define wordCount as Multiset:

for each token in T{

wordCount[token]++; }

for each document in documentSet { T=tokenize(document);





第 5 章 MapReduce 分布编程模型

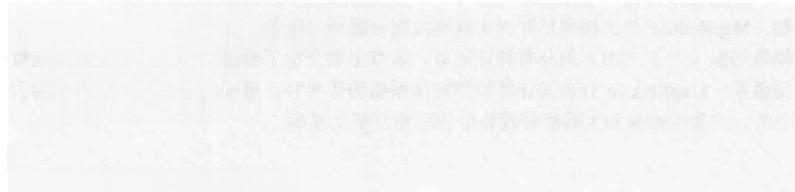


}

Display(wordCount);

该程序循环遍历所有的文档。对于每个文档，使用分词过程逐个地提取单词。对于每 个单词，在多重集合 wordCount 中的相应项上加1。最后， display(函数打印出 wordCount 中的所有条目。

这个程序只适合处理少量文档， 一旦文档数量激增，它就不能胜任了。例如，需要编 写一个垃圾邮件过滤器，来获取接收到的几百万封垃圾邮件中经常使用的单词。使用单台 计算机反复遍历所有文档将会非常费时。重写程序，让工作可以分布在多台机器上。每台 机器处理这些文档的不同部分。当所有的机器都完成时，第二个处理阶段将合并这些结果。 第一阶段要分布到多台机器上去的伪代码为：



define wordCount as Multiset;

for each document in documentSubset(

T=tokenize(document);

for each token in T(

wordCount[token]++;

}

}

sendTOSecondphase(wordCount);

第二阶段的伪代码为：

Define totalWordCount as Multiset;

for each wordCount received from firstPhase{

multisetAdd(totalWordCount,wordCount);

}

首先，忽略了文档读取的性能需求。如果文件都存在一个中央存储服务器上，那么瓶

颈就是该服务器的带宽。让更多的机器参与处理的办法将不会一直有效，因为有时存储服 务器的性能跟不上。因此，需要将文档分开存放，使每台机器可以仅处理自己所存储的文 档，从而消除单个中央存储服务器的瓶颈。这就是说，在数据密集型分布式应用中存储和 处理不得不紧密地绑定在一起。

该程序的另一个缺陷是 wordCount 和 totalWordCount 被存储在内存中。当处理大型文 档集时， 一个特定单词的数量就会超过一台机器的内存容量。英语有大约一百万个单词， 这个大小可以很轻松地放进一个iPod, 但单词统计程序将处理许多不在标准英语单词词典 中的特殊单词。例如，必须处理诸如 Hadoop 这样的特定名称。我们必须统计错字，即使 它们并不是真正的单词(例如， exampel), 我们还需分别统计一个字的所有不同形式(例 如 ，eat,ate,eaten 和 eating) 。 即使在文档中特定单词的数量可以被管理在内存中，在问 题的定义上略有变化就可能会引起空间复杂度的爆炸。例如，不统计文档中的单词，而是 去统计日志文件中的IP 地址，那么我们处理的多重集合条目将达到数十亿，这超过了大多 数计算机的内存容量。

此外，第二阶段只有一台计算机，它将处理来自所有计算机在第一阶段计算 wordCount 的结果。wordCount 的处理任务原本就相当繁重。当我们为 Map 的处理提供充足的计算机



**大数据基础与应用**

时 ，Reduce 阶段的单台计算机将成为瓶颈。最明显的问题是，能否按分布模式重写Reduce 阶段，以便它可以通过增加更多的计算机来实现扩展。

为了使 Reduce 阶段以分布的方式运行，必须以某种方式将其分割到多台计算机上执 行，使之能够独立运行。需要在第一阶段之后将 wordCount 分区，使得 Reduce 阶段的每台 计算机仅需处理一个分区。例如，假设在 Reduce 阶段有26 台计算机。每台计算机上的 wordCount 只处理以特定字母开头的单词。计算机A 在 Reduce 阶段仅统计以字母a 开头的 单词，以此类推。

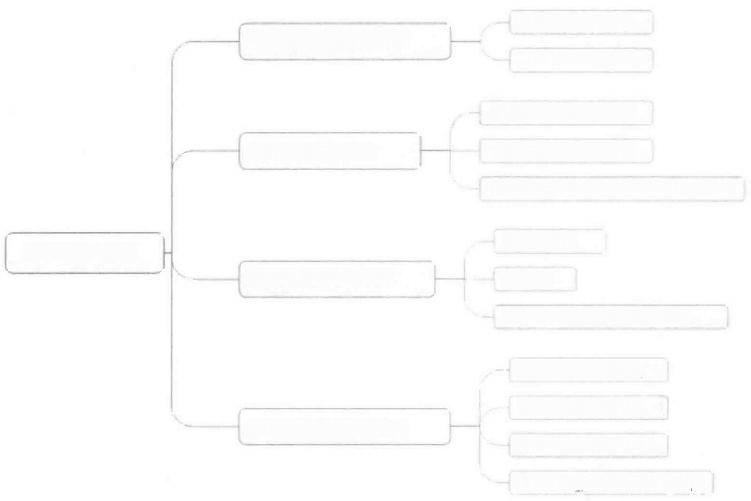
小结

算法是计算机科学的基石。任何一个计算问题的分析与建模，几乎都可以归为算法问 题。MapReduce 算法模型是针对大规模数据处理而提出的分布编程模型，主要应用于大规 模数据集(大于1TB) 的分布并行运算。本章主要介绍了函数式编程范式、映射函数与化 简函数、MapReduce 分布式计算模型的体系结构等内容。通过本章的学习，可以为大数据 分析、大数据挖掘和大数据可视化分析等建立坚实基础。



**第6章** **大数据流式计算**

本章主要内容



流式数据的概念

流式数据的概念与特征

流式数据的特征

大数据批量计算模型

大数据流式计算模型

大数据流式计算与批量计算的比较

Storm系统

大数据流式计算

S4系统

Data Freeway and Puma系统

金融银行业的应用

互联网领域的应用

大数据流式计算的应用

物联网领域的应用

三种典型应用场景的对比

流式大数据处理工具

大数据的计算模式

大数据的存储、处理、安全及应用都已成为研究的热点。尤其是大数据包括了越来越 多不同格式的数据，从简单的电子邮件、数据日志和信用卡记录，到仪器收集到的科学大 数据、医疗大数据、财务大数据以及丰富的媒体数据，包括照片、音乐和视频等。这些不 同格式的数据需要使用不同的处理方法。上述问题导致应用批量数据处理技术处理大数据 出现了性能瓶颈。为此，流式数据处理技术和交互式数据处理技术应运而生，发展迅速。

**6.1** **流式数据的概念与特征**

数据密集型应用的特征是：不宜用持久稳定关系建模，而适宜用瞬态数据流建模。典型 应用的实例包括金融服务、网络监控、电信数据管理、Web 应用、生产制造、传感检测等。

**6.1.1** **流式数据的概念**

流式数据是指产生的数据不是批量地传输过来，而是像水一样“流”过来。流式数据

**大数据基础与应用**

的处理也是连续处理，而不是批量处理。如果是全部收到数据以后再以批量的形式处理， 那么延迟时间很长，而且在很多场合将消耗大量内存资源。

**6.1.2** **流式数据的特征**

**1.** **实时性**

由于数据源的种类多而复杂，导致了数据流中的数据可以是结构化的数据、半结构化 的数据，甚至是无结构化的数据。数据源不受任何接收系统的控制，所以数据的产生是实 时的、连续不断的、不可预知的。也就是说，流式大数据是实时产生、实时计算，计算结 果的反馈也往往需要保证及时性。流式大数据的大部分数据到来之后直接在内存中进行计 算，并在计算之后被丢弃，只有少量数据长久地保存到硬盘中，这就需要系统计算快，计 算延迟足够小。在数据价值有效的时间内，体现数据的有用性，因此，可以优先计算时效 性特别短、潜在价值又很大的数据。

**2.** **易失性**

通常数据流到达后立即计算并使用，只有极少数的数据需持久地保存下来，大多数数 据直接丢弃。数据的使用往往是一次性的、易失的。即使重放，得到的数据流与得到之前 的数据流往往也是不同的，这就需要系统具有一定的容错能力，能够充分地利用仅有的一 次数据计算的机会，尽可能全面、准确、有效地从数据流中获得有价值的信息。

**3.** **突发性**

数据的产生完全由数据源确定，由于不同的数据源在不同时空范围内的状态不统一， 导致数据流的速率呈现出了突发性变化的特征。前一时刻数据速率和后一时刻数据速率可 能有巨大的差异，数据的流速波动较大，这就需要系统具有很好的可伸缩性，能够动态适 应不确定流入的数据流，具有很强的系统计算能力和大数据流量动态匹配的能力。进而达 到在高数据流速的情况下不丢弃数据，也可以识别并选择丢弃部分不重要的数据，在低数 据速率的情况下，保证不长时间占用系统资源。

**4.** **无序性**

大数据的无序性是指各数据流之间无序，而同一数据流内部各数据元素之间也无序。 其原因如下：

① 由于各个数据源之间是相互独立的，所处的时空环境也不尽相同，因此无法保证各 数据流间的各个数据元素的相对顺序。

② 即使是同一个数据流，由于时间和环境的动态变化，也无法保证重放数据流和之前 数据流中数据元素顺序的一致性。这就需要系统在数据计算过程中具有很好的数据分析和 发现规律的能力，不能仅依赖数据流间的内在逻辑或者数据流内部的内在逻辑。

③ 流式数据的元组通常带有时间标签或含序属性。因此，同一流式数据往往是被按序 处理，然而数据的到达顺序不可预知，由于时间和环境的动态变化，无法保证重放数据流





第 6 章 大数据流式计算

与之前数据流中数据元素顺序的一致性，进而导致了数据的物理顺序与逻辑顺序不一致， 即数据流顺序颠倒，或者由于丢失而不完整。

**5.** **无限性**

数据实时产生并动态增加，只要数据源处于活动状态，数据就一直产生和持续增加下 去。潜在的数据量无法用一个具体确定的数据描述，在数据计算过程中，系统无法保存全 部数据。这是由于既没有足够大的硬件空间来存储无限增长的数据，也没有合适的软件来 有效地管理这么多数据，更无法保证系统长期而稳定地运行。

**6.** **准确性**

数据的质量不能保证就是准确性不能保证。在大数据中，将重复数据、异常数据和不 完整数据统称为脏数据，由于数据流中含有脏数据不可避免，因此流式数据的处理系统需 要对脏数据具有很强大的数据抽取和动态清洗等能力，进而获得高质量的数据。

**6.2** **大数据的计算模式**

大数据的计算模式可以分为批量计算(batch computing)和流式计算(stream computing) 两种基本形态。

**6.2.1** **大数据批量计算模型**

大数据批量计算模式如图6-1 所示，批量计算首先将 数据存储到硬盘中，然后再对存储在硬盘中的静态数据进 行集中计算。Hadoop 是典型的大数据批量计算架构，由 HDFS 分布式文件系统负责静态数据的存储，并通过 MapReduce 将计算逻辑分配到各数据节点进行数据计算 和价值发现。

**6.2.2** **大数据流式计算模型**

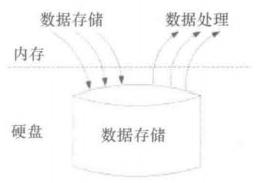


图6-1 大数据批量计算

**1.** **数据流管理系统**

数据以一个或多个流的方式到来， 一方面，如果不对数据进行及时的处理和存储，数 据将永远丢失。另一方面，数据的传输速度太快，致使不可能将全部数据存储于数据库中 并在选定的时间进行交互式处理。

流式数据计算的算法都在某种程度上包含流的汇总过程。首先需要考虑从流中获取样 本的方法，以及从流中滤除部分不需要的元素。然后展示估计流中的独立元素个数，估计 方法所用的开销远少于列举所有所见元素的开销。另一种对流进行汇总的方法是只观察一 个定长窗口，该窗口由最近的n 个元素组成，其中n 为某个给定值，然后将这个窗口当做

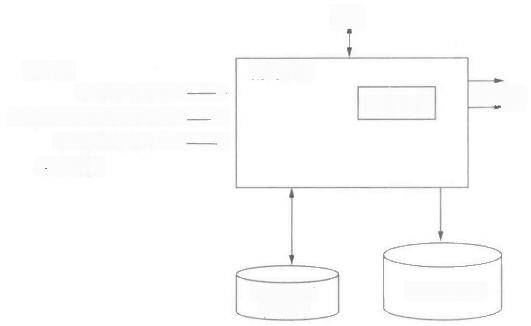




**大数据基础与应用**

一个数据库一样进行查询处理。如果有很多流或者n 很大，可能无法存储每个流的整个窗 口。这时需要对这些窗口进行汇总处理。

可以将数据流处理器看成一种数据管理系统，如图6-2所示。



查询

输入流

3,4,5,6,7,8,9—→

A,S,D,F,G,H,H,J,K,L →

1,0.0,1,1,<1.0.0.1> →

时间

有限工作 存储器

固定查询 输出流

归档存储器

流处理器

图6-2 数据流管理系统

若干数量的流进入系统，每个流可以按照各自的时间表提供元素，各个流的数据率和 数据类型不相同，流中的元素到达的时间间隔不一定需要满足均匀分布。流元素的到达速 率也不受系统控制，由于数据库管理系统控制数据从磁盘读出的速率，不易产生数据丢失， 所以流处理与数据库中的数据处理不同。可以将数据流在大容量的归档存储器中归档处理， 虽然不能再归档存储器中进行应答查询，但可以将流汇总数据或者部分流数据存于工作存 储器中。在工作存储器中可以完成查询应答处理，存储器容量有限，可以是磁盘或内存， 取决于查询处理的速度要求。

流查询有两种方式， 一种是固定查询，另一种是窗口查询。

(1)固定查询

永远不变地执行查询并在适当时刻产生输出结果。例如，查询一个数据流，当超过某 个值时就发出警报，由于该查询仅依赖于最近的那个流元素，因此对其进行处理相当容易。

(2)ad hoc 查询

这种查询对于当前某个或者多个数据流仅提供一次，如果没有存储所有的流数据，那 么系统就不能应答关于流的任意查询。完成这种查询任务有多种查询方法， 一种通用的方 法是在工作存储器上保存每个流的滑动窗口。 一个滑动窗口是最近到达的n 个流元素，也 可以是在最近 t 个时间单位中到达的所有元素。如果将每个流元素看成一个元组，那么就 可以将窗口看成是关系数据库，在其上执行任意的 SQL 查询。当然，流管理系统必须在新 元素到达时删除最早的那些元素，进而保持窗口的最新值。

**2.** **大数据流式计算**

大数据流式计算模型如图6-3所示，流式计算中，无法确定数据的到来时刻和到来顺 序，也无法将全部数据存储起来。因此，不再进行流式数据的存储，而是当流动的数据到 来后在内存中直接进行数据的实时计算。如 Twitter 的 Storm 、Yahoo 的 S4(simple scalable





**第6章** **大数据流式计算**

streaming system) 就是典型的流式数据计算架构，数据在内存中被计算，并输出有价值的 信息。

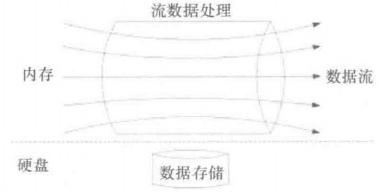


图6-3 大数据流式计算

**6.2.3** **大数据流式计算与批量计算的比较**

流式计算、批量计算分别适用于不同的大数据应用场景。对于先存储后计算、实时性 要求不高，但对数据的准确性和全面性更为重要的应用场景，批量计算模式更为合适。对 于无须先存储，可以直接进行数据计算，实时性要求很严格，但数据的精确度要求稍微宽 松的应用场景，更适于流式计算。在流式计算中，由于数据在最近一个时间窗口内，所以 数据延迟往往较短，实时性较强，但数据的精确程度较低。流式计算和批量计算具有互补 特征，在多种应用场合下可以将两者结合起来，通过发挥流式计算的实时性优势和批量计 算的计算精确性优势，来满足多种应用场景的数据计算。

大数据批量计算相关技术的研究相对成熟，形成了 MapReduce 编程模型、开源的 Hadoop 计算系统为代表的高效、稳定的批量计算系统，在理论上和实践中取得了显著成果。 关于流式计算的早期研究集中在数据库环境中开展数据计算的流式化，数据规模较小，数 据对象比较单一。由于大数据呈现出实时性、易失性、突发性、无序性、无限性等特征， 对系统提出了更高的要求。2010年Yahoo 推出S4 流式计算系统，2011年，Twitter 推出 Storm 流式计算系统，在一定程度上推动了大数据流式计算技术的发展和应用。但是，这些系统 在可伸缩性、系统容错、状态一致性、负载均衡、数据吞吐量等多方面仍然存在着明显不 足。构建低延迟、高吞吐量、持续可靠运行的大数据流式计算系统平台，是当前亟待解决 的问题。

大数据流式计算与批量计算在各个主要性能指标上的比较结果如表6-1 所示。

**表6-1** **大数据流式计算与批量计算的比较**

|  |  |  |
| --- | --- | --- |
| **性** **能** **指** **标** | **大数据流式计算** | **大数据批量计算** |
| 计算方式 | 实时 | 批量 |
| 常驻空间 | 内存 | 硬盘 |
| 时效性 | 短 | 长 |
| 有序性 | 无 | 有 |



大数据基础与应用

续表

|  |  |  |
| --- | --- | --- |
| **性** **能** **指** **标** | **大数据流式计算** | **大数据批量计算** |
| 数据量 | 无限 | 有限 |
| 数据速率 | 突发 | 平稳 |
| 是否可重现 | 难 | 稳定 |
| 移动对象 | 数据移动 | 程序移动 |
| 数据精确度 | 较低 | 较高 |

**6.3** **流式大数据处理工具**

流式大数据处理已经逐渐得到广泛的应用，目前典型的大数据流式计算系统有：

● Twitter 的 Storm 系统；

● Yahoo 的 S4 系统；

● Facebook 的 Data Freeway and Puma 系统；

● Linkedin 的 Kafka 系统；

● Microsoft 的 TimeStream 系统；

● Hadoop 之上的数据分析系统 HStreaming;

● IBM 的商业流式计算系统 StreamBase;

● Berkeley 的交互式实时计算系统 Spark;

● 复杂事件处理的 Esper系统等。

下面介绍比较典型的、应用较为广泛的、具有代表性的前3种大数据流式计算系统进 行实例分析。

**6.3.1** **Storm** **系统**

Storm 是 Twitter 开发的一种分布式的、开源的、实时的、主从式大数据流式计算系统， 版本是 Storm0.8.2, 使用的协议为Eclipse Public License 1.0,其核心部分用高效流式计算 的函数式语言Clojure 编写，极大地提高了系统性能。但为了方便用户使用，支持用户使用 任意编程语言进行项目的开发。

**1.** **任务拓扑**

任务拓扑是Storm 的逻辑单元， 一个实时应用的计算任务将被打包为任务拓扑后发布， 任务拓扑一旦提交后将一直运行，除非显式地中止。 一个任务拓扑是由一系列 Spout 和 Bolt 构成的有向无环图，通过数据流实现 Spout 和 Bolt 之间的关联，如图6-4所示。其中，Spout 负责从外部数据源不间断地读取数据，并以元组的形式发送给相应的 Bolt,Bolt 对接收到 的数据流进行计算，即实现过滤、聚合、查询等具体功能，可以级联，也可以向外发送数 据流。

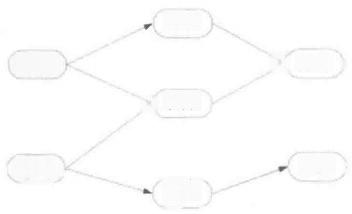




第 6 章 大数据流式计算

数据流是时间上无穷的元组序列，如图6-5 所示。数据流通过流分组所提供的不同策 略实现在任务拓扑中流动。此外，为了满足确保消息能且仅能被计算1 次的需求，Storm

还提供了事务任务拓扑。



Bolt

 Bolt

7 Bolt

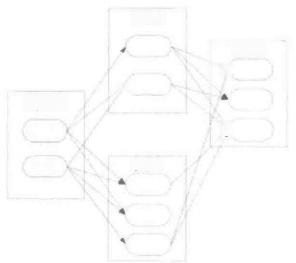
Spout

Bolt

Spout

Bolt

图6-4 Storm任务拓扑

Bolt

Bolt

2



Spout

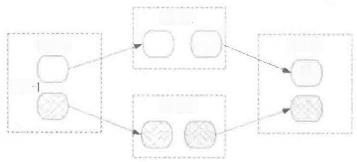
4

Bolt

图6-5 Storm 数据流组

**2.** **作业级容错机制**

用户可以为一个或多个数据流作业(以下简称数据流)进行编号，分配一个唯一的ID, Storm 可以保证每个编号的数据流在任务拓扑中完全执行。完全执行是指由该ID 绑定的源 数据流以及由该源数据流后续生成的新数据流，应该到达的 Bolt 被完全执行。如图6-6所 示，两个数据流被分配ID=1, 当且仅当两个数据流分别经过Bolt 1和 Bolt 2, 最终都到达 Bolt 3 并都被完全处理后，才表明数据流被完全执行。

Bolt I

Bolt 3

Spout

ID=

Bolt 2

图6-6 数据流作业完全执行

Storm 通过系统级组件 Acker 实现对数据流的全局计算路径的跟踪，并保证该数据流被 完全执行。其基本原理是为数据流中的每个分组进行编号，并通过异或运算来实现对其计 算路径的跟踪。

作业级容错的基本原理如下：

Axor A=0,

A xor B…xor B xor A=0,当且仅当每个编号仅出现2次。

作业级容错的基本流程是：在 Spout 中，系统会为数据流的每个分组生成一个唯一的 64位整数，作为该分组的根ID 。根 ID 会被传递给 Acker 及后续的 Bolt 作为该分组单元的 唯一标识符。同时，无论是 Spout 还是 Bolt, 每次新生成一个分组的时候，都会重新赋予 该分组一个新的64位的整数的ID 。Spout 发送完某个数据流对应的源分组后，并告知Acker 自己所发射分组的根ID 及生成的那些分组的新 ID, 而 Bolt 每次接收到一个输入分组并计



Slaves Cluster Master

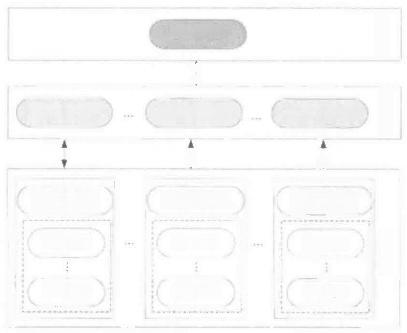


**大数据基础与应用**

算完之后，也将告知 Acker 自己计算的输入分组的ID 及新生成的那些分组的ID 。Acker 只 需要对这些ID 做一个简单的异或运算，就能判断出该根ID 对应的消息单元是否完成计算。

**3.** **总体架档**

Storm 采用主从系统架构，如图6-7所示，在一个 Storm 系统中有两类节点，即一个主 节点 Nimbus、多个从节点 Supervisor 。Storm有3种运行环境，即 Master、Cluster 和 Slaves。



Nimbus



\*

Zookeeper

T

Supervisor

Worker

Worker

Supervisor

Worker

Supervisor

Worker

Zookeeper

Zookeeper

Worker

Worker

T

图6-7 Storm系统架构

(1)主节点Nimbus

主节点运行在 Master 环境中，是无状态的，负责全局的资源分配、任务调度、状态监 控和故障检测。

① 主节点 Nimbus 接收客户端提交来的任务，验证后分配任务到从节点 Supervisor 上 ， 同时把该任务的元信息写入 Zookeeper 目录中。

② 主节点 Nimbus 需要通过 Zookeeper 实时监控任务的执行情况，当出现故障时进行 故障检测，并重启失败的从节点 Supervisor 和工作进程 Worker。

(2)从节点 Supervisor

从节点运行在 Slaves 环境中，也是无状态的，负责监听并接收来自主节点 Nimbus 所 分配的任务，并启动或停止自己所管理的工作进程 Worker, 其中，工作进程 Worker 负责 具体任务的执行。 一个完整的任务拓扑往往由分布在多个从节点 Supervisor 上的 Worker 进 程来协调执行，每个 Worker都执行且仅执行任务拓扑中的一个子集。在每个 Worker 内部， 会有多个Executor, 每个 Executor 对应一个线程。Task 负责具体数据的计算，即，用户所 实现的 Spout/Blot 实例。每个 Executor 会对应一个或多个 Task, 因此，系统中 Executor 的 数量总是小于等于Task 的数量。

(3)Zookeeper

Zookeeper 是一个针对大型分布式系统的可靠协调服务和元数据存储系统，通过配置 Zookeeper 集群，可以使用 Zookeeper 系统所提供的高可靠性服务。Storm 系统引入 Zookeeper, 极大地简化了Nimbus 、Supervisor 、Worker 之间的设计，保障了系统的稳定性。





第 6 章 大数据流式计算

Zookeeper 在 Storm 系统中具体实现了以下功能。

●存储客户端提交的任务拓扑信息、任务分配信息、任务的执行状态信息等，便于 主节点Nimbus 监控任务的执行情况。

● 存储从节点 Supervisor、工作进程 Worker 的状态和心跳信息，便于主节点Nimbus 监控系统各节点运行状态。

● 存储整个集群的所有状态信息和配置信息，便于主节点 Nimbus 监控 Zookeeper 集群的状态，在主Zookeeper 节点挂掉后，可以重新选取一个节点作为主 Zookeeper 节点，并进行恢复。

**4.** **系统特征**

Storm 系统的主要特征包括：

(1)编程模型简单

用户只需编写 Spout 和 Bolt 部分的实现程序，因此极大地降低了实时大数据流式计算 的复杂性。

(2)支持多种编程语言

默认支持 Clojure 、Java 、Ruby 和 Python, 也可以通过添加相关协议实现对新增语言的 支持。

(3)作业级容错性

可以保证每个数据流作业被完全执行。

(4)水平可扩展

计算可以在多个线程、进程和服务器之间并发执行。

(5)快速消息计算

通过ZeroMQ 作为其底层消息队列，保证了消息能够得到快速的计算。

**5.** **不足之处**

① 资源分配没有考虑任务拓扑的结构特征，无法适应数据负载的动态变化。 ② 采用集中式的作业级容错机制，在一定程度上限制了系统的可扩展性。

**6.3.2** **S4** **系统**

S4 是 Yahoo 支持开发的一种分布式的、可扩展的、可插拔的、对称的大数据流式计算 系统，使用的协议为 Apache License 2.0, 编程语言为 Java。

**1.** **处理单元** **PE**

处理单元 PE(Processing Element) 组成如图6-8所示，由下述4个组件构成。

● 函数：实现了与该处理单元PE 相对应的功能和配置。

● 事件类型：规定了该处理单元 PE 所接收的事件类型。

● 主键：规定了该处理单元PE 所关心的事件主键。

● 键值：规定了该处理单元 PE 所匹配的键值。



6 1



**大数据基础与应用**

处理单元 PE 只关心与其事件类型相匹配的事件，并仅处理与其主键、键值相 一 致的 事件，即只有事件类型、主键、键值全部匹配之后，处理单元 PE 才会处理该类事件。当 一 个新事件没有可以匹配的处理单元PE 时，系统将会为该事件新创建 一 个处理单元 PE。

因此，不仅需要高效、动态地创建、管理和删除处理单元PE, 还需要更合理地规划处理单 元 PE 的类型设计及其拓扑结构。

应说明的是，有 一 类处理单元 PE 位 于S4 的输入层，它们没有主键、键值，只需事件 类型相匹配，即对该类事件进行处理。通常情况下，该类处理单元 PE 所计算的事件为原 始输入事件，其输出事件会被新增主键、键值替代，以便后续处理单元 PE 进行计算。

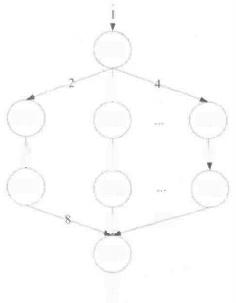
**2.** **任务拓扑结构**

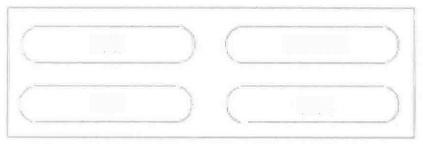
在 S4 系统中，数据流是由事件的有序序列 (K,A) 构 成 的 ， 其 中 ，K,A 分 别 表 示 该

*类型事件的若干个* *key 和若干个* *attribute,key 和* *attribute 都是* *tuple-valued, 即* *key=value*

的元组值。事件在各个处理单元 PE 中被计算，在处理单元 PE 之间流动，处理单元 PE 之 间的逻辑构成了 一 个有向无环图。

在图6 - 9所示的有向无环图中，节点表示处理单元 PE, 实现对数据流的计算和新数据 流的输出，有向边表示事件的有序序列 (K,A) 及其流向。

PEl

事件类型

函数

主键

键值

图6-8 处理单元 PE 组成

PE2

5

I

PE5

图6-9

3

PE3

PE4

7

PE7)

PE6

3

PE8

10

I1

\*

任务拓扑实例

**3.** **处理节点** **Pnode**

如图6-10所示， S4 的处理节点 Pnode 由处理空间和传输空间两部分组成。 ( 1 ) 处 理 空 间

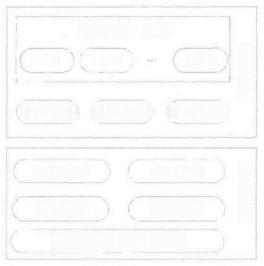
在处理空间中，事件监听系统主要用于监听并分发接收到的事件计算请求，并由调度 分配系统将事件分配到处理单元集 (Processing Element Container,PEC) 上进行计算。处 理 单 元 集 PEC 以适当的顺序调用适当的处理单元 PE, 并保证每个主键 key 的处理单元 PE 都会被映射到 一 个确定的处理节点 Pnode 上。之后，处理节点 Pnode 或者发出输出事件， 或者向传输层请求协助，向指定逻辑节点发送消息。其中，处理单元集PEC 由 一 个处理节 点 Pnode 中内部的多个处理单元PE 组成。处理单元 PE 是事件计算的最小单元，接受 一 个





**第6章** **大数据流式计算**

或多个来自事件源或其他处理单元 PE 的事件进行计算，在此之后分发一个或多个计算后 的事件到其他处理单元 PE 或输出结果。各个处理单元 PE 间相互独立无关，它们之间通过 事件构成关联，事件在各处理单元PE 间以数据流的形式进行传输。



处理单元集PEC

PE 2

事件监听) (调度分配) (事件分发

路由管理 负载均衡

集群管理 容错管理

可靠协调系统Zookeeper

传输空同

处理空间

PE 1

PEn

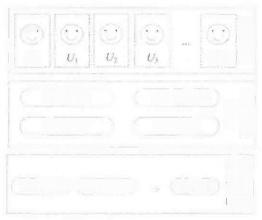
图6-10 处理节点 Pnode

(2)传输空间

在传输空间中，主要通过路由管理、负载均衡、集群管理、容错管理等实现对事件流 的路由选择、负载均衡、逻辑影射、故障恢复到备用节点等方面的管理和功能，并通过 Zookeeper 系统在 S4 集群节点间实现一致性协作。S4 通过插件式的架构来动态选择信息传 输协议，对于控制信息，通常采用可靠传输协议，如 TCP, 保障控制信息传输的可靠性。 对于数据信息，通常采用不可靠传输协议，如 UDP, 保障数据信息的高吞吐量。

**4.** **系统架构**

S4 采用了对等式系统架构，如图6-11 所示。





U₀

客户适配器

名字眼务

Pnode 1 Pnode 2

性能监控

配置维护



思 空 同

资 源 两 度

Pnode n

高

U

图6-11 S4系统结构

一个 S4 系统由用户空间、资源调度空间和 S4 处理节点空间组成。其中，在用户空间 中，多个用户可以通过本地的客户端驱动实现服务的请求访问。在资源调度空间中，为用 户提供了客户适配器，通过TCP/IP 协议实现用户的客户端驱动与客户适配器间的连接和通 信，多个用户可以并发地与多个客户适配器进行服务请求。在 S4 处理节点空间中，提供了 多个处理节点 Pnode, 进行用户服务请求的计算。各个处理节点间保持相对的独立性、对 等性和高并发性，极大地提高了系统的性能，并通过 Hash 方式将事件传输到一个或多个目





**大数据基础与应用**

标处理节点 Pnode 上。

**5.** **不足之处**

S4 系统存在的不足：

① 当数据流到达速度超过一定界限时，到达速度越快，系统数据处理的错误率越大。 ② 不支持系统节点的热插拔，所有对节点的调整都必须离线进行。

③ 仅支持部分容错，即节点失效转移时会丢失原节点内存中的状态信息。

**6.3.3** **Data** **Freeway** **and** **Puma** **系统**

Data Freeway and Puma 是 Facebook 支持开发的一种基于 Hive/Hadoop 的、分布式的、 高效率的、数据传输通道和大数据流式计算系统。

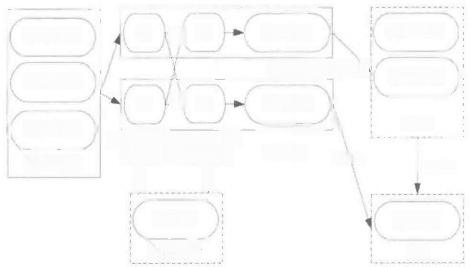
**1.Data Freeway 系统**

Data Freeway 是 Facebook 支持开发的一种可扩展数据流架构，可以有效地支持4种数 据间的传输，即文件到文件、文件到消息、消息到消息和消息到文件。其系统结构如图6-12 所示，数据流架构由下述4个组件构成。

① PTail.Scribe 组件位于用户端，其功能是将用户的数据通过 RPC 发送到服务器端。

② Calligraphus 组件实现了对日志类型的维护与管理，其功能是通过 Zookeeper 系统， 将位于缓冲区中的数据并发写到HDFS 中。

③ Continuous Copier 组件的功能是实现在各个 HDFS 系统间进行文件的迁移。 ④ PTail 组件实现了并行地将文件输出。



HDFS节点

HDFS节点

数据节点

HDFS

Calligraphus Calligraphus

HDFS层 PTail

PTail

Log节点

Log

Seribe节点

Scribe节点

Scnbe节点

Seribe组件

数据节点

Continuous copier

ZK节点

Zookeeper

C₁2 C₂

写数据层

中间层

A T



4 中

C₁

C₂

图6-12 Data Freeway 系统架构

**2.Puma 系统**

Puma 是 Facebook 的可靠数据流聚合引擎 (reliable stream aggregation engine) 系统， 如图6-13所示。



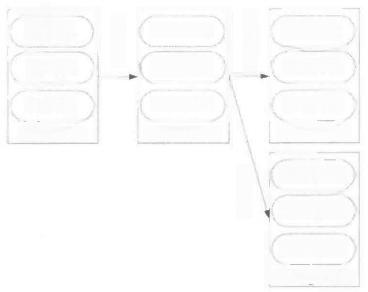
各份数据流

读数据流

毁 据 流



第6章 大数据流式计算

HBase节点

Ptail节点

Ptail节点

Piail节点

PTail子系统

Puma3节点

Puma3节点

Puma3节点

Puma3 子系统

HBase节点

HBase节点

HBase子系统

Serving节点

Serving节点

Serving 节点

Serving子系统

图6-13 Puma3 系统架构

Puma 在本地内存中实现了数据聚合功能，提高了数据的计算能力，有效地降低了系统 延迟。Puma 系统实现时，在 Calligraphus 组件通过聚合主键完成对数据的分片，其中，每 个分片都是内存中的哈希表，每个表项对应一个Key 及用户定义的聚合方法，如统计、求 和、平均值等操作。HBase 子系统会定期地将内存中的数据备份到HBase 中，进行数据的 持久化存储。当Puma 发生故障时，从HBase 中读取副本，实现对因故障丢失数据的恢复。 在无故障的情况下， HBase 子系统不参与数据的计算，因此提高了数据的计算能力。

**3.** **不足之处**

Data Freeway and Puma 系统存在的不足：

① 数据延迟在秒级，无法满足大数据流式计算所需要的毫秒级应用需求。 ② 将哈希表完全放入内存的加速机制，导致内存需求量大。

③ 资源调度策略不够简单、高效，不能灵活适应连续的工作负载。

**6.4** **大数据流式计算的应用**

大数据流式计算模型主要适用于商业计算与科学计算中的快速运算和分析数据，也适 于Web 网站、博客、电子邮件、视频、新闻、电话记录、传输数据、电子感应器之中的具 有一定格式的数据流进行处理的需求。大数据流式计算主要用于对动态产生的数据进行实 时计算并及时反馈结果，但不要求结果绝对精确。在数据的有效时间内获取其价值，是大 数据流式计算系统的首要设计目标，因此，当数据到来后将立即对其进行计算，而不再对 其进行缓存等待后续全部数据到来再进行计算。

大数据流式计算的应用场景较多，按照数据产生方式、数据规模以及技术成熟度这3 个维度考虑，选择金融银行业应用、互联网应用和物联网应用这3种典型应用场景，分析 与说明大数据流式计算的基本特征。



**大数据基础与应用**

**6.4.1** **金融银行业的应用**

在金融银行领域的日常运营过程中产生大量数据，由于这些数据的时效性较短，因此， 金融银行领域是大数据流式计算最典型的应用场景之一，也是大数据流式计算最早的应用 领域。在金融银行系统内部，每时每刻都有大量的结构化的数据在各个系统间流动，并需 要实时计算。同时，金融银行系统与其他系统也有着大量的数据流动，这些数据不仅有结 构化数据，也会有半结构化和非结构化数据。通过对这些大数据的流式计算，发现隐含于 其中的内在特征，可以帮助金融银行系统进行实时决策。

在金融银行的实时监控场景中，大数据流式计算的优势如下。

**1.** **风险管理**

风险管理包括信用卡诈骗、保险诈骗、证券交易诈骗、程序交易等，需要实时跟踪 发现。

**2.** **营销管理**

根据客户信用卡消费记录，掌握客户的消费习惯和偏好，预测客户未来的消费需求， 并为其推荐个性化的金融产品和服务。

**3.** **商业智能**

掌握金融银行系统内部各系统的实时数据，实现对全局状态的监控和优化，并提供决 策支持。在金融银行领域的日常运营过程中将产生时效性较短，不仅有结构化数据，也会 有半结构化和非结构化数据。通过对这些大数据的流式计算，发现隐含于其中的内在特征， 可帮助金融银行进行实时决策。传统的商业智能要求数据是静态的，通过数据挖掘技术， 获得数据的价值。然而在瞬息万变的场景下，如股票期货市场，需要数据挖掘技术能及时 地响应需求，这就需要应用流式数据处理技术。

**6.4.2** **互联网领域的应用**

随着互联网技术的不断发展，特别是Web 2.0 时代的到来，用户可以实时分享和提供 各类数据。不仅使得数据量大为增加，也使得数据更多地以半结构化和非结构化的形态呈 现。据统计，目前互联网中75%的数据来源于个人，主要以图片、音频、视频数据形式存 在，需要实时分析和计算这些大量的动态数据。

**1.** **搜索引擎**

在互联网领域中，搜索引擎提供商们将在反馈给客户的搜索页面中加入点击付费的广 告信息。插入什么广告、在什么位置插入这些广告才能得到最佳效果，需要根据客户的查 询偏好、浏览历史、地理位置等综合语义进行决定。而这种计算对于搜索服务器来说是大 量的计算。主要表现在每时每刻都会有大量客户进行搜索请求，而且由于数据计算的时效





第 6 章 大数据流式计算

性极低，需要保证极短的响应时间。

**2.** **社交网站**

社交网站需要实时分析用户的状态信息，及时把用户的最新信息分享给相关的朋友。 准确地推荐朋友，推荐主题，提升用户体验，并能及时发现和屏蔽各种欺骗行为。

**6.4.3** **物联网领域的应用**

在物联网环境中，各个传感器产生大量数据。这些数据通常包含时间、位置、环境和 行为等内容，具有明显的颗粒性。由于传感器的多元化、差异化以及环境的多样化，这些 数据呈现出鲜明的异构性、多样性、非结构化、有噪声、高增长率等特征。所产生的数据 量之密集、实时性之强、价值密度之低，需要进行实时、高效的计算。实时分析提供动态 的信息展示，目前主要应用于智能交通、环境监控、灾难预警等。Web 数据采集系统是利 用网络爬虫程序抓取万维网上的内容，通过清洗、归类、分析并挖掘其数据价值。在物联 网领域中，大数据流式计算的典型应用场景如下。

**1.** **智能交通**

通过传感器实时感知车辆、道路的状态，并分析和预测一定范围、 一段时间内的道路 流量情况，以便有效地进行分流、调度和指挥。

**2.** **环境监控**

通过传感器和移动终端，对一个地区的环境综合指标进行实时监控、远程查看、智能 联动、远程控制，系统地解决综合环境问题。这些对计算系统的实时性、吞吐量、可靠性 等方面都提出了很高要求。

**6.4.4** **三种典型应用场景的对比**

大数据流式计算的三种典型应用场景的对比如表6-2所示。

**表6-2** **三种典型应用场景的比较说明**

|  |  |  |  |
| --- | --- | --- | --- |
| **典** **型** **应** **用** | **数据产生方式** | **数** **据** **规** **模** | **技术成熟度** |
| 金融银行业 | 被动产生数据 | 小规模 | 成熟度高 |
| 互联网 | 主动产生数据 | 中规模 | 成熟度中 |
| 物联网 | 自动产生数据 | 大规模 | 成熟度低 |

**1.** **数据的产生方式**

从数据的产生方式看，金融银行领域的数据是在系统中被动产生的，互联网领域的数 据是人为主动产生的，物联网领域的数据往往是由传感器等设备自动产生的。

**2.** **数据的规模**

从数据的规模来看，金融银行领域的数据与互联网、物联网领域的数据相比较少；物





**大数据基础与应用**

联网领域的数据规模是最大的，但受制于物联网的发展阶段，当前实际拥有数据规模最大 的是互联网领域。

**3.** **技术成熟度**

从技术成熟度来看，金融银行领域的流式大数据应用最为成熟，从早期的复杂事件处 理开始就呈现了大数据流式计算的思想。互联网领域的发展，将大数据流式计算真正推向 历史舞台，物联网领域的发展为大数据流式计算提供了重要的历史机遇。

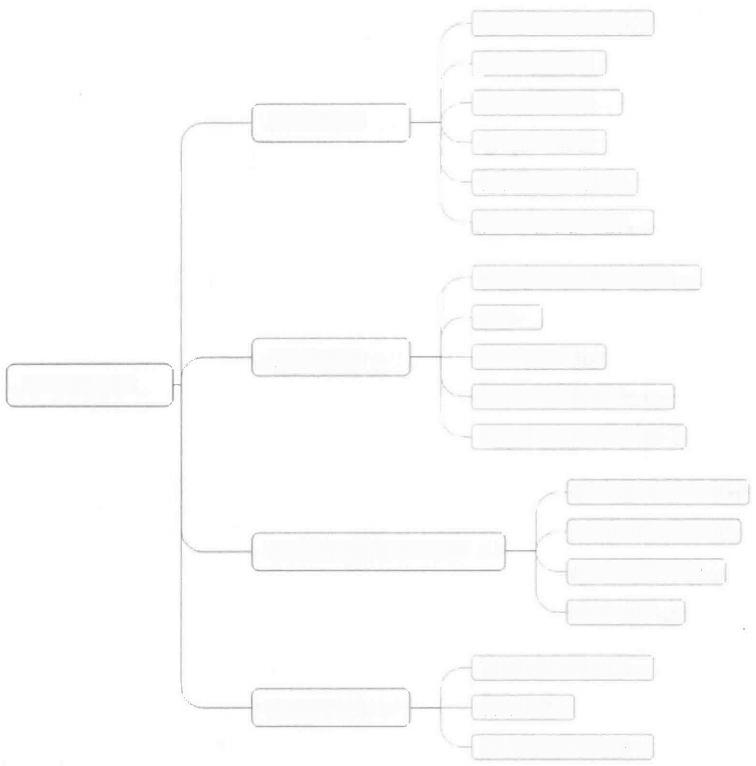
**小结**

流式大数据处理技术在商业智能、市场营销和公共服务等诸多领域有着广泛的应用前 景，并已在金融银行业、互联网、物联网等场景的应用中取得了显著的成效。流式大数据 以其实时性、无序性、无限性、易失性、突发性等显著特征，与传统批量大数据在数据计 算的要求、方式等方面有着明显的不同，也使得当前数据计算系统无法进一步更好地适应 流式大数据在系统可伸缩性、容错、状态一致性、负载均衡、数据吞吐量等方面所带来的 新挑战。本章介绍了流式数据特征和流式大数据处理技术，分析和对比了当前较为典型的 三种大数据流式计算系统。



**第** **7** **章** **大数据搜索技术**

本章主要内容



搜索引擎的发展过程

搜索引擎的定义

搜索引擎的组成

搜索引擎概述

搜索引擎的分类

搜索引擎的工作过程

搜索引擎的评价指标

语义与语义搜索引擎的概念

图谱

搜索就是回答

大数据搜索技术

语义搜索引擎的组成

基于本体的语义搜索引擎

垂直网站与垂直搜索

私有化的Web化数据

网站数据对搜索引擎的影响

没有Web化的数据

大数据流动

搜索引擎优化的产生

网页级别

搜索引擎优化的方法

语义搜索引擎

搜索引擎优化

网络信息检索现在已经成为获取信息的主要手段，统计结果表明，目前中国用户上网 的最主要目的是获取信息，占42.3%,位居榜首。有98.7%的用户通过互联网获取信息， 其中有71.9%的用户通过搜索引擎来查找相关网站。面对数量如此巨大的用户，网络信息 检索面临下述两个最为突出、亟待解决的问题：

● 搜索的结果相关度低，冗余信息太多；

● 智能水平不高，搜索引擎无法回答常识性问题。

**大数据基础与应用**

出现上述问题的主要原因如下。

一是检索技术主要依据编码技术，通过分类模式描述给定的信息，然后使用字符串匹 配的全文检索方法搜索用户所提交的关键词。由于使用这种编码技术，描述的信息只能反 映出部分语义，因此不能保证语义的完全匹配。

二是检索过程是把用户的查询关键词与全文中的每个词进行遍历式比较，而没有考虑 查询请求与文档语义上的匹配。

针对上述问题，众多学者进行了研究，尤其是大数据出现之后，基于语义搜索引擎模 型与方法的研究已成为热点并被寄予厚望。应用语义搜索技术，可以将全世界范围内的数 据通过链接进入图谱，整合成一个巨大的数据库，搜索引擎就变成了一个问答引擎，能够 理解前后语境进而回答问题。

**7.1** **搜索引擎概述**

**7.1.1** **搜索引擎的发展过程**

互联网出现之后，流行网站分类目录查询。由人工整理维护网站分类目录，主要工作 是精选互联网上的优秀网站，并简要描述，分类存放于不同目录下。当用户查询时，通过 逐层点击来查找所需要的网站。当时将基于目录检索服务的这种网站称为搜索引擎。

1993年，出现了在网站上使用的搜索软件，通过分析字词关系，对互联网上的大量信 息做更为有效的检索。

1995年，出现了元搜索引擎。元搜索引擎的工作方式是：用户只需提交一次搜索请求， 由元搜索引擎负责转换处理后提交给多个预先选定的独立搜索引擎，并将从各独立搜索引 擎返回的所有查询结果集中起来处理，之后再返回给用户。

之后，出现了智能检索。智能检索利用分词词典、同义词典、同音词典来改善检索效 果，进一步还可在知识层或者概念层上辅助查询。通过主题词典、上下位词典、相关同级 词典来检索处理形成一个知识体系或概念网络，给予用户智能知识提示，最终帮助用户获 得最佳的检索效果。

例如查询“计算机”,那么与计算机相关的信息也能检索出来，并可以进一步缩小查询 范围至微机、服务器或扩大查询至信息技术或查询相关的电子技术、软件和计算机应用等， 包括歧义信息和检索处理。再如查询“苹果”,究竟是指水果还是计算机品牌，将通过歧义 知识描述库、全文索引、用户检索上下文分析以及用户相关性反馈等技术结合处理，高效、 准确地反馈给用户最需要的信息。

个性化搜索引擎能够通过社区化产品(即对注册用户提供服务)的方式来组织个人信 息，然后在搜索引擎基础信息库的检索中引入个人因素进行分析，获得针对个人的不同搜 索结果。分析特定用户的搜索需要限定的范围，然后按照用户需求范围扩展到互联网上其 他的同类网站给出的最相关结果。

由于没有统一的信息组织标准对网络信息资源进行加工处理，难以对无序的网络信息



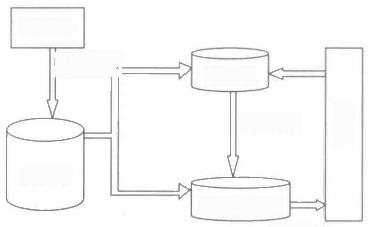


第7章 大数据搜索技术

资源进行检索、交接和共享乃至深层次的开发利用，形成了信息孤岛。利用互联网可以消 除信息孤岛，并实现互联网上资源的全面连通。

**7.1.2** **搜索引擎的定义**

搜索引擎 (Search Engine) 是指根据一定的策略、运用特定的计算机程序，从互联网 上搜集信息，在对信息进行组织和处理之后为用户提供检索服务，将用户检索的相关信息 展示给用户的系统，如图7-1所示。



Internet

抓取数据

读取记录

资料库

有序资料库

显示结果

资料库

用 户

图7-1 搜索引擎

**7.1.3** **搜索引擎的组成**

搜索引擎由搜索器、索引器、检索器和用户接口四个部分组成。搜索器的功能是在互 联网中漫游，发现和搜集信息。索引器的功能是理解搜索器所搜索的信息，从中抽取出索 引项，用于表示文档以及生成文档库的索引表。检索器的功能是根据用户的查询在索引库 中快速检出文档，进行文档与查询的相关度评价，对将要输出的结果进行排序，并实现某 种用户相关性反馈机制。用户接口的作用是输入用户查询、显示查询结果、提供用户相关 性反馈机制。

**7.1.4** **搜索引擎的分类**

搜索引擎主要分为下述几类。

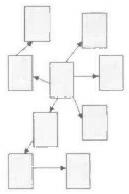
**1.** **全文搜索引擎**

全文搜索引擎是指从网站上提取信息来建立网页数据库。全文搜索引擎的自动信息搜 集功能分为定期搜索和提交网站搜索两种。

(1)定期搜索

定期搜索是指每隔一段时间(例如Google 一般是28天)进行一次搜索。蜘蛛搜索引 擎可以定期主动执行“蜘蛛”程序，对一定IP 地址范围内的互联网网站进行检索， 一旦发 现新的网站，就自动提取网站的信息和网址加入自己的数据库。





G

A

E

H I

D

C

B

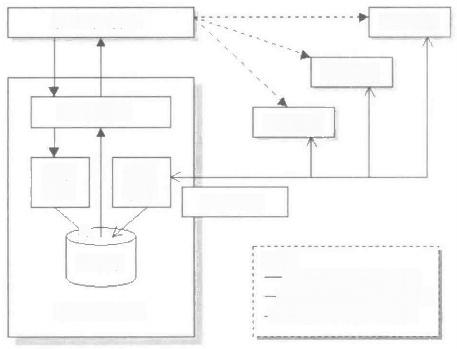
F



**大数据基础与应用**

(2)提交网站搜索

提交网站搜索是指网站拥有者主动向搜索引擎提交网址，它在一定时间内(2天到数月 不等)派出抓取程序，扫描网站并将有关信息存入数据库，以备用户查询。随着搜索引擎索 引规则发生很大变化，主动提交网址并不保证网站能进入搜索引擎数据库。最好的办法是多 获得一些外部链接，让搜索引擎有更多机会找到网站，并自动将网站收录，如图7-2所示。



Web站点

Web站点

Web服务器

Web站点

检索

软件

搜索软件

x

图中：

→表示数据库的更新

→表示检索的查询过程 ---→表示浏览Web页面

数据库

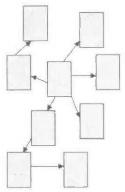
搜索引擎主机

客户端浏览器

索引 软件

图7- 2 提交网站搜索

抓取搜索引擎的搜索方式分为广度优先和深度优先两种搜索顺序，如图7-3 所示。当 用户以关键词查找信息时，搜索引擎会在数据库中进行搜寻，如果找到与用户要求内容相 符的网站，便根据网页中关键词的匹配程度、出现的位置、频次、链接质量计算出各网页 的相关度及排名等级，然后根据关联度高低，按顺序将这些网页链接返回给用户。这种引 擎的特点是搜全率比较高。



B

F

A

E

H

G

D

C

I

广度优先的抓取顺序： 深度优先的抓取顺序：

A—BCDEF—HG— 1 A—F—G

E—H—1

图7- 3 抓取搜索引擎的搜索方式

**2.** **目** **录** **索** **引** **引** **擎**

( 1 ) 目 录 索 引 引 擎 的 概 念

目录索引也称为分类检索，是因特网上最早提供 WWW 资源查询的服务，主要通过搜



**第7章** **大数据搜索技术**

集和整理因特网的资源，根据搜索到的网页内容，将其网址分配到相关分类主题目录的不 同层次、类目之下，形成像图书馆目录一样的分类树形结构索引。目录索引无须输入任何 文字，只要根据网站提供的主题分类目录，逐层点击进入，便可查到所需的网络信息资源。

虽然目录索引具有搜索功能，但严格意义上不能称为真正的搜索引擎，只是按目录分 类的网站链接列表而已。用户完全可以按照分类目录找到所需要的信息，不依靠关键词进 行查询。

(2)目录索引引擎与全文搜索引擎的比较

全文搜索引擎与目录索引引擎具有下述不同之处。

① 全文搜索引擎是自动进行网站检索，而目录索引引擎则完全依赖手工操作，用户提 交网站后，目录编辑人员将亲自浏览网站，然后根据一套自定的评判标准决定是否接纳需 检索的网站。其次，只要网站本身没有违反有关的规则，通过全文搜索引擎一般都能登录 成功。目录索引引擎对网站的要求则高得多，有时即使登录多次也不一定成功。尤其像 Yahoo 这样的超级索引，登录更是困难。

② 全文搜索在登录搜索引擎时， 一般不考虑网站的分类问题，而登录目录索引时则必 须考虑将网站放在一个最合适的目录。

③ 全文搜索引擎中各网站的有关信息都是从用户网页中自动提取的，所以用户拥有更 多的自主权，而目录索引引擎则要求必须手工填写网站信息，而且还有各种各样的限制。 更有甚者，如果工作人员认为提交网站的目录、网站信息不合适，可以随时对其进行调整， 当然事先不与用户商量。

全文搜索引擎与目录索引引擎有相互融合渗透的趋势。 一些纯粹的全文搜索引擎也提 供目录搜索。

**3.** **元搜索引擎**

元搜索引擎接受用户查询请求后，同时在多个搜索引擎上搜索，并将结果返回给用户。 著名的元搜索引擎有InfoSpace 、Dogpile 、Vivisimo 等，中文元搜索引擎中具代表性的是搜 星搜索引擎。在搜索结果排列方面，有的直接按来源排列搜索结果，如 Dogpile; 有的则按 自定的规则将结果重新排列组合，如 Vivisimo。

**4.** **垂直搜索引擎**

垂直搜索是针对某一个行业的专业搜索引擎，是搜索引擎的细分和延伸，是对网页库 中的某类专门的信息进行一次整合，定向分字段抽取出需要的数据进行处理后再以某种形 式返回给用户。与垂直搜索相对应的通用搜索引擎信息量大、查询不准确、深度不够等。 提出来的新的搜索引擎服务模式，能够通过针对某一特定领域、某一特定人群或某一特定 需求提供有一定价值的信息和相关服务。其特点就是专、精、深，且具有行业色彩，相比 较通用搜索引擎的海量信息无序化，垂直搜索引擎则显得更加专业、具体和深入。

垂直搜索引擎为逐步兴起的一类搜索引擎。不同于通用的网页搜索引擎，垂直搜索专 注于特定的搜索领域和搜索需求(如机票搜索、旅游搜索、生活搜索、小说搜索、视频搜 索、购物搜索等),在其特定的搜索领域有更好的用户体验。垂直搜索硬件成本低、用户需 求特定、查询的方式多样。



**大数据基础与应用**

**5.** **集合式搜索引擎**

集合式搜索引擎类似于元搜索引擎，区别在于它并非同时调用多个搜索引擎进行搜索， 而是由用户从提供的若干搜索引擎中选择。

**6.** **门户搜索引擎**

门户搜索引擎虽然提供搜索服务，但自身既没有分类目录也没有网页数据库，其搜索 结果完全来自其他搜索引擎，如AOL Search 、MSN Search等。

**7.1.5** **搜索引擎的工作过程**

搜索引擎的工作过程经历了爬行、抓取存储、预处理、排名四个步骤，分述如下。

**1.** **爬行**

搜索引擎是通过一种特定的软件跟踪网页的链接，从一个链接爬到另外一个链接，像 蜘蛛在蜘蛛网上爬行一样，所以被称为蜘蛛，也被称为机器人。

**2.** **抓取存储**

搜索引擎是通过蜘蛛跟踪链接爬行到网页，并将抓取的数据存入原始页面数据库。其 中的页面数据与用户浏览器得到的完全一样。搜索引擎蜘蛛在抓取页面时，也做了重复内容 检测， 一旦遇到权重很低的网站上有大量抄袭、采集或者复制的内容，很可能就不再爬行。

**3.** **预处理**

搜索引擎将蜘蛛抓取回来的页面，进行下述的预处理。

① 提取文字；

② 中文分词；

③ 去停止词；

④ 识别并消除噪声(如版权声明文字、导航条、广告等);

⑤ 正向索引；

⑥ 倒排索引；

⑦ 链接关系计算；

⑧ 特殊文件处理。

除 了HTML 文件外，搜索引擎通常还能抓取和索引以文字为基础的多种文件类型，如 PDF 、Word 、WPS 、XLS 、PPT 、TXT 文件等，但搜索引擎还不能处理图片、视频、Flash 类非文字内容，也不能执行脚本程序。

**4.** **排名**

用户在搜索框中输入关键词后，排名程序调用索引库数据，计算排名显示给用户，排 名过程与用户直接互动。但是，由于搜索引擎的数据量庞大，虽然能达到每日都有小的更 新，但是一般情况下，搜索引擎的排名规则是根据日、周、月阶段性更新。



第7章 大数据搜索技术

**7.1.6** **搜索引擎的评价指标**

**1.** **查全率**

查全率又称为召回率，是指用户通过搜索引擎所获得的有用信息与整个Internet 中相关 信息的比率。

**2.** **查准率**

查准率是指用户通过搜索引擎所获得的真正需要的信息占获取信息的比率。

**3.** **响应时间**

响应时间是指用户发出查询请求之后至看到查询结果的这段时间。

**4.** **覆盖范围**

覆盖范围是指搜索引擎索引的Web 页面占整个 Internet 中页面的比例。

**5.** **用户方便性**

用户方便性主要包括查询接口是否直观、是否易于使用，查询语法是否丰富，显示结 果是否易于查看等。

**7.2** **语义搜索引擎**

**7.2.1** **语义与语义搜索引擎的概念**

**1.** **语义**

语义是指语言的意义或者与语言的意义相关的文本。语义技术研究问题集中于信息含 义的获得，并利用数据分析、数据挖掘和可视化分析的支持工具，获取深度内涵和推论。 语义搜索技术是许多重要技术研究的综合结晶。

**2.** **语义搜索引擎**

利用语义搜索引擎可以自动识别文本的概念结构，并通过上下文来解读搜索结果。例 如，如果搜索“计算机软件”一词，语义搜索引擎可能获取软件工程、软件构造、软件体 系结构和软件测试等文本信息。如果计算机软件这个词没有出现在信息来源中，语义搜索 可以对关键词的相关词和类似词进行解读，从而扩大搜索信息的准确性和相关性。

在语义搜索的过程中，需要对查询的信息和网络上的内容进行歧义消除，即搜索引擎 通过自然语言处理消除模棱两可的情况。例如搜索“美洲虎”(捷豹轿车),就会知道要找 一辆车还是一个大型猫科动物。目前，搜索引擎基本上是根据关键词的关注度进行排序， 将搜索频度高的关键词结果列在前面。



**大数据基础与应用**

算法与人的思维主要的区别是人可以直接理解词或文章的意思，但算法却不能理解。 人看到“苹果”这两个字后，脑中就会浮现出水果的形象，这是由于苹果的概念在人脑中 已经形成，但是，搜索引擎却不能从感性上理解。由于搜索引擎可以掌握词之间的关系， 这就牵涉语义分析，也就是通过海量文献寻找出词汇之间的关系。当两个词或一组词大量 出现在同一个文档中时，就可以认为这些词之间语义相关。例如，电脑和计算机这两个词 在人们写文章时经常混用，这两个词在大量的网页中同时出现，搜索引擎就认为这两个词 是语义相关。

语义索引并不依赖于语言，所以 SEO 和搜索引擎优化虽然一个是英语， 一个是中文， 但这两个词大量出现在相同的网页中，即使搜索引擎并不知道搜索引擎优化或 SEO 指的是 什么,但是却可以从语义上把“SEO”“搜索引擎优化”“search engine optimization""SEM” 等词紧紧地联系在一起。

(1)网站主题的形成

网站分成不同栏目，各栏目介绍有区别但紧密相关的话题，这些话题共同形成了网站 主题。搜索引擎在把整个网站的页面收录进去之后，能够根据这些主题词之间的语义相关 程度，判断出网站的主题。

(2)网页内容写作

搜索引擎排名有一个现象，即搜索某个关键词时，排在前面的网页有时甚至并不含有 所搜索的关键词，这有可能是潜在语义索引在起作用。例如搜索“电脑”,排在前面的网页 有可能只提到计算机却没提到电脑。因为搜索引擎通过语义分析知道这两个词是紧密相关 的。还有一个要注意的是，在进行网页写作的时候，不要局限于目标关键词，应该包含与 主关键词语义相关、相近的词汇，以支持主关键词。这在搜索结果中也有体现，有的文章 虽然大量出现主关键词，但缺少其他支撑词汇，搜索排名并不理想。

网络上大部分的搜索问题都是随机漫步式的浏览，对结果的要求也不精益求精，所以 语义搜索的用处不大，对于具有针对性的研究搜索，语义搜索作用巨大。

**7.2.2** **图谱**

随着大数据的兴起，从大量无结构数据中获取上下文是一个困难的问题。但是利用图 谱数据库，语义技术可以将每个人和事物联系起来，可以保证高质量的搜索结果。知识图 谱是大数据语义链接的基石，也称为科学知识图谱，它将应用数学、图形学、信息可视化 技术、信息科学等学科的理论方法与计量学引文分析、共现分析等方法结合，并利用可视 化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构，达到多学 科融合的目的，为学科研究提供切实的、有价值的参考。知识图谱实现对客观世界从字符 串描述到结构化语义描述，是对客观世界的知识映射。知识图谱可以描述不同层次和粒度 的概念抽象，可以作为互联网资源组织的基础。虽然语义 Web 的愿景还尚未发生，知识图 谱的发展是让互联网更好地具有世界知识的良好开端。在网络搜索时，经常会出现多义的 词条，如具有相同名字的几个人，通常搜索结果将以结果列表的形式给出。知识图谱的语 义链接，使得搜索引擎可以用基于实体的搜索来代替基于字符串的搜索，从而实现搜索时





**第** **7** **章** **大数据搜索技术**

的歧义消除。知识图谱使互联网从字符串描述发展到客观世界的具体事物描述，互联网为 知识图谱构建提供了丰富的资源，知识图谱是互联网理解世界的基础设施。

Google 利用 Knowledge Graph系统描述了人、地点、物体间的相互关系。现在，这样 的语义网络包含了5.7亿个体，超过180亿个关于不同个体的相互关系，这一切都将用于 更好地理解用户输入的关键词，了解用户的倾向和意图。当一个用户提出疑问，语义技术 便会为用户生成一个信息图谱，从不同角度为原先的提问形成全面的概括。这种工具从用 户的原始搜索出发，将不同的搜索结果有机结合。同样，Facebook 也推出了一个连接不同 用户的社交图谱，将搜索逐渐转型为寻找，使得搜索更具有智能性。

**7.2.3** **搜索就是回答**

基于语义和图谱搜索现已实现。当大数据与图谱出现后，语义技术将打破传统的商务 市场，以便留住用户，扩大规模。当消费者通过搜索技术寻求复杂问题的答案时，利用语 义技术，市场人员可以快速向用户推销他们之前寻找过的商品。Facebook 、Google 、IBM 还有微软都早已看到了语义技术的未来和价值。现在，搜索就是回答。

**7.2.4** **语义搜索引擎的组成**

语义搜索引擎由开放云、数据工厂和大脑三个部分组成。

**1.** **开放云**

开放云主要面向开发者，引擎的开放云则是面向有数据存储和处理需求的开发者。例如， 百度的开放云拥有超过1.2万台的单集群，还拥有CPU 利用率高、弹性高、成本低等特点。

**2.** **数据工厂**

如果将开放云理解为基础设施和硬件能力，则可以把数据工厂理解为将海量数据组织 起来的软件能力，就像数据库软件的地位一样，只不过数据工厂被用作处理 TB 级甚至更 大的数据。

**3.** **大脑**

图灵奖获得者沃斯 (N.Wirth) 提出“程序=数据结构+算法”的理论。将语义搜索引擎 看作是一个程序，那么数据结构是“数据工厂+开放云”,而算法则对应大脑。可以将语义 搜索引擎描述为“语义搜索引擎=数据工厂+开放云+大脑”。

**7.2.5** **基于本体的语义搜索引擎**

基于本体的语义搜索引擎是运用本体论构建基于本体的语义搜索引擎模型。该模型能 够根据用户的查询关键字或者询问问题，进行基于知识的推理，从而提高检索结果的相关 度，进而实现较高水平的语义检索。





**大数据基础与应用**

**1.** **本体的定义**

本体这个术语源于哲学，本体在人工智能领域得到广泛的应用，但目前还没有形成标 准定义，仅将常用的三种定义列举如下。

(1)概念模型的形式化说明

本体是对共享概念模型的形式化说明，主要表现为下述几点。

① 概念模型：指通过抽象客观世界中一些现象的相关概念而得到的模型，其表示的含 义与具体环境无关。

② 明确的定义：指所使用的概念及使用这些概念的约束定义明确。 ③ 形式化：表现在本体可读。

④ 共享：指本体中的共同认可的知识，反映的是相关领域中公认的概念集，所针对的 是团体，而不是个体。简单地说，本体给出构成相关领域词汇的基本术语和关系，以及利 用这些术语和关系构成的确定词汇外延的有关规则的定义。其目标是捕获相关领域的知识， 提供对该领域知识的共同理解，确定领域内通用的词汇，并给出这些词汇之间相互关系的 明确定义。

(2)词汇或概念的理论

本体是关于词汇或概念的理论，本体是一种表示性的词汇，可以应用于特定领域。例 如，计算机科学与技术领域的本体包含一些描述基本概念的词汇：计算机系统、运算器、 软件等，也包含这些基本词汇间的关系。例如，计算机是计算机系统的一部分，而运算器 是计算机的一个部件元件。

(3)定义领域知识主体

本体用来定义某一领域的知识主体。本体是描述某个领域的知识，它不仅仅是简单的 词汇表，而是整个上层知识库，包括用于描述这个知识库的词汇。本体可以作为知识图谱 表示的概念模型和逻辑基础。

综上所述，本体是某个领域内(可以是特定专业的，也可以是宽泛范畴的)不同主体 之间进行交流的一种语义基础，即由本体提供定义明确的词汇表，描述概念和概念之间的 关系，作为使用者之间达成的共识。

**2.** **本体的作用**

(1)本体可为人与人之间或组织与组织之间的通信提供交流的共同词汇。 (2)本体可以实现不同系统之间集成。

(3)本体可以重用，有助于知识获取，提高系统可靠性和规范描述。

**3.** **本体语义搜索引擎**

(1)基于本体搜索引擎的设计

本体提供了人机交流的机制，从而为搜索引擎效率提高奠定了基础。基于本体的搜索 引擎的基本思想如下所述。

① 在领域专家的帮助下，建立相关领域的本体。

② 收集信息源中的数据，并参照已建立的本体，把收集来的数据按规定的格式存储在





第7章 大数据搜索技术

元数据库中。

③ 查询转换器把查询请求转换成规定的格式，在本体的帮助下从元数据库中找出符合 条件的数据集合。

④ 检索的结果经过定制处理后，返回给用户。

(2)知识库

知识库是语义搜索引擎进行推理和存储知识积累的容器，而本体则是知识库的基础。 本体提供一组术语和概念来描述某个领域，知识库则使用本体表达该领域的事实。本体和 知识库的关系为：

① 本体为知识库的建立提供一个基本的结构。

② 本体提供一套概念和术语来描述某一领域，并且获取该领域的本质的概念结构。 ③ 知识库运用这些术语去表达现实或者虚拟世界中的正确知识。

综上所述，可以看出构建知识库的第一步就是对该领域进行有效的本体分析。

**4.** **构造本体**

本体是基于本体的信息检索系统的基石，决定系统的性能、通用程度以及系统运行的 质量。如何正确、有效、合乎逻辑地建立本体是系统建立的关键。

(1)本体的构造准则

对同一领域、同一事物，不同人可能建立不同的本体。由于本体应该是规范化的描述， 因此遵循统一的构造准则。目前最为常用的是格鲁伯(Gruber) 提出的构造本体的5条准则：

① 清晰：本体必须能够有效地说明所定义的术语。

② 一致：本体应该能够支持与其定义相一致的推理。

③ 可扩展性：可以支持在已有的概念基础上定义新的术语。

④ 编码偏好程度最小：概念的描述不依赖于某一种特殊的符号层的表示方法。 ⑤ 本体约定最小：只要能够满足特定的知识共享需求即可。

(2)本体的表示

本体表示方法应用广泛。 一是传统的四元素表示方法，二是较新的六元组表示法。前 者过于灵活，不易掌握；后者因为定义规范，可操作性强。

① 四元素表示方法： 一个本体含有概念、关系、实例和公理四个主要元素。其中概念 表示某个领域中一类实体或事物的集合；关系描述概念和概念属性的交互；实例是概念表示 的具体的事物；公理用来限制类和实例的取值范围，在公理中包括许多具体的规则和约束。

② 六元组表示法：六元组表示法用一个六元组来表示一个本体，即

An Ontology={C,AC,R,AR,H,X}

其中，C 表示概念的集合； AC 表示多个属性集合组成的集合，其中每个属性集合对应 于一个概念； R 是一个关系集合； AR 是由多个属性集合组成的集合，其中每个属性集合对 应于R 中的一个关系； H 表示概念之间的层次结构关系； X 表示公理集合。

(3)本体构造的生命周期

首先明确本体使用的目的和范围，然后依次构造本体。其具体过程可以分为下述三个 主要阶段。





**大数据基础与应用**

① 本体捕获阶段：确定关键的概念和关系，给出精确定义，并确定其他相关的术语。

② 本体编码阶段：选择描述语言来表达概念和术语。已存本体的集成是对已存本体的 重用和修改，这个阶段也是一个循环往复的迭代过程。

③ 评估阶段：根据需求描述、能力问题等对本体以及软件环境、相关文档进行评价。

**5.** **基于本体的语义搜索引擎模型简介**

基于本体的语义搜索引擎模型是基于本体的搜索引擎，能够实现语义搜索、知识检索 和推理功能。本模型搜索引擎的环境是 Web 网页，这些网页并未自动含有语义标记。此系 统也应具备搜索引擎的各种基本功能，例如网页的遍历和获取、索引的建立、页面查找算 法等，都可以参考目前流行的搜索引擎的结构和实现过程。

在基于本体的语义搜索引擎模型中，信息库和知识库可以交流。知识库是实现智能搜 索的核心，知识库的丰富也决定了检索能力和回答问题能力的高低。智能搜索引擎就是通 过知识库把用户的问题提高到知识层面，然后利用知识检索信息库。二者的有机结合离不 开语义分析和知识管理。

**7.3** **网站数据对搜索引擎的影响**

聚合所有网络上的信息是搜索引擎所追寻的目标，1994 年，吉尔·艾尔斯沃斯博士 (Dr.Jill Ellsworth) 提出暗网的概念，暗网是指不能被标准搜索引擎索引的网络。研究结 果表明，人类信息只有不到1%实现了Web 化，而Web 化的网页中，搜索引擎能抓取的大 概为1%。

影响搜索引擎抓取的因素既有网站本身非主观的问题，如不符合网页规范、对搜索引 擎不友好等，但也有网站本身的主观屏蔽的问题，如淘宝、优酷等网站屏蔽百度的爬虫。

**7.3.1** **垂直网站与垂直搜索**

**1.** **垂直网站**

垂直网站集中于某些特定的领域或某种特定的需求而建立，提供关于这个领域或需求 的全部深度信息和相关服务，垂直网站越来越受重视并得以广泛应用。

**2.** **垂直搜索引擎**

垂直搜索引擎是针对某一个行业的专业搜索引擎，是搜索引擎的垂直延伸，是对网页 库中的某类专门的信息进行一次整合，定向分字段抽取出需要的数据进行处理后再以某种 形式返回给用户。

网页搜索是以网页为最小单位，而垂直搜索是以结构化数据为最小单位。垂直搜索引 擎将网页的非结构化数据抽取成特定格式的结构化数据，然后将这些数据存储到数据库， 进行进一步的处理，例如除掉重复数据、分类等，最后分词、索引。经过上述过程之后，





**第** **7** **章** **大数据搜索技术**

将非结构化数据抽取成了结构化数据，再经过深度加工处理后返回给用户。

垂直搜索引擎的应用广泛，例如企业库搜索、供求信息搜索引擎、购物搜索、房产搜 索、人才搜索、地图搜索、MP3 搜索、图片搜索等，几乎各行各业各类信息都可以进一步 细化成各类的垂直搜索引擎。

购物搜索引擎的流程如下：抓取网页后，对网页商品信息进行抽取，抽取出商品名称、 价格、简介等，如可以将笔记本电脑简介进一步细分成品牌、型号、CPU、 内存、硬盘、 显示屏……然后对数据进行清洗、去重、分类、分析比较、挖掘，最后通过分词索引提供 用户搜索，通过分析挖掘提供市场行情的报告。

垂直搜索引擎需要蜘蛛、网页结构化信息抽取技术或元数据采集技术、分词、索引等其 他信息处理技术。垂直搜索引擎的技术评估应从全面性、更新性、准确性和功能性来判断。

**3.** **垂直搜索引擎的特点**

● 垂直搜索引擎抓取的数据来源于垂直搜索引擎关注的行业站点；

● 垂直搜索引擎抓取的数据倾向于结构化数据和元数据；

● 垂直搜索引擎的搜索行为是基于结构化数据和元数据的结构化搜索。

**7.3.2** **私有化的** **Web化数据**

垂直网站在达到一定规模后，拥有与搜索引擎博弈的能力，可以屏蔽搜索引擎的搜索， 将自己的数据私有化。电商网站、BBS、 互动百科等内容属于私有化的 Web 数据。对于垂 直网站来说，可以用个性化的搜索功能和独有的挖掘能力，完成更好的搜索，甚至可上升 为垂直搜索引擎。另外一种垂直搜索引擎即是综合其他垂直的结构化数据，提供搜索服务。

垂直搜索是搜索引擎的一个发展方向，将对传统搜索引擎构成挑战。手机上的浏览器 和App 的流量对半分。将传统搜索引擎比喻为浏览器，那么垂直搜索引擎便是App 。垂直 搜索引擎正在迅速壮大，使 Web 数据中的不能抓取私有化信息所占比例变得更小。

**7.3.3** **没有** **Web 化的数据**

移动互联网已积累大量的数据，App 、云应用、社交和物联网使数据爆炸式增长。但 是，这些数据对搜索引擎透明。

**1.** **人工整理的数据**

没有通过网站公开的大量民生数据已达到可观的量级。

**2.** **社交产生的数据**

部分社交网站的数据是 Web 化的，但是它们是封闭的。搜索引擎对它们无能为力。

**3.** **个人云应用产生的数据**

个人云应用将越来越多，数据量越来越大，但搜索引擎对这些数据无能为力。



**大数据基础与应用**

**4.** **物联网产生的数据**

车联网、监控录像、电子抄表、水文监测等物联网应用产生了大量的数据，搜索这些 数据需要考虑搜索形式和搜索结果等问题。

**7.3.4** **大数据流动**

将数据私有化或者有限地开放给部分搜索引擎，这就涉及大数据流动问题。

**1.** **远离搜索引擎的数据**

垂直网站将其数据私有化，社交网站本身就将数据私有化，而云应用提供商保存着用 户的私有数据，App 的数据因为没有Web 化所以也是私有化的，当然还有一部分数据掌握 在政府和普通企业手中。这些数据都是远离搜索引擎的数据。

现在结构化的数据，尤其是有价值的结构化数据正在远离搜索引擎，流向一个私有的 区域。这将产生数据的滚雪球效应：有数据的地方，数据会越来越多；没有数据的地方， 必须为获得数据付出比蜘蛛爬取更多的代价。

**2.** **搜索引擎将退化**

传统综合搜索引擎需要解决的问题不是加速信息流动，而是解决找到信息的问题。这 也更加突出收集数据对搜索引擎的重要性。传统的网页搜索引擎将退化为垂直网页搜索引 擎。因为网页数据只是网络数据的一部分。当然，还有一种可能是搜索引擎仍然可以找到 这些数据，但是需要有偿获取。

**3.** **大数据对搜索价值重大**

按照达尔文生物进化论，人类的信息吸收、筛选和处理的能力应该也会进化。人们对 信息的需求并不会退化，反而会更加饥渴。而搜索引擎需要解决的问题，不再是帮助人们 从海量信息里面找到结果，而是在海量结果中找到唯一答案。

(1)结构化数据对搜索引擎的价值

结构化数据和网页数据相比，能够找准唯一答案。网页分析是靠文本匹配。结构化数 据的分析既支持内容提供者的主动接入，也支持搜索引擎的个性化精准分析。这两种方式 都会增加内容提供者或者搜索引擎的成本，但是付出带来的回报是用户快速得到准确的、 唯一的答案。

(2)大数据挖掘需要搜索引擎

搜索引擎需要帮助人类完成人脑不能做的事情——数据挖掘，即从海量数据中挖掘价 值。搜索引擎经过十多年的发展，在文本分析、关系发掘、图谱构造、用户语义理解等方 面已有丰富的积累。这些技术是大数据挖掘依赖的基本技术，又可称为挖掘引擎。

**7.4** **搜索引擎优化**

搜索引擎优化 (Search Engine Optimization,SEO) 是指通过研究搜索引擎的运作规则



**第7章** **大数据搜索技术**

来调整网站，进而提高网站在搜索引擎的排名的方式。由于搜索引擎的用户通常只注意搜 索结果最前面的几个条目，所以不少网站都希望通过各种形式来影响搜索引擎的排序，使 自己的网站可以获得理想的搜索排名结果。搜索引擎优化的处理就是为了使网站更容易被 搜索引擎接受。搜索引擎将网站彼此间的内容做相关性的数据比较，然后再由浏览器将这 些内容以最快速、最完整的方式呈现给搜索者。搜索引擎优化是指通过搜索引擎的规则进 行优化，为用户打造更好的用户体验，其最终目的就是做好用户体验。对于任何一个网站 来说，要想在网站推广中取得成功，搜索引擎优化都是至为关键的一项任务。同时，随着 搜索引擎不断变换它们的搜索排名算法规则，每次算法上的改变都将使网站排名发生变化， 都将在网站中引起反响。可以说，搜索引擎优化是一个越来越复杂的任务。

**7.4.1** **搜索引擎优化的产生**

搜索引擎优化始于20世纪90年代中期。当时所有网站管理员的工作是提交所属网站 到各个搜索引擎。这些引擎利用一些蜘蛛机器人，截取网页程序中连接到其他网页的超链 接，并且存储所找到的数据。这个过程中同时包含了将网页下载并存储至搜索引擎拥有者 的服务器中，利用Indexer 软件来截取页面中各种信息，主要包括页面中的文字、文字的位 置、文字的重要性以及页面所包含的任何链接，此后再将页面置入清单中等待， 一段时间 之后，再来截取一次。

搜索引擎碰到许多滥用与操纵排名等问题。为了给使用者提供较好的结果，搜索引擎 必须调整搜索结果表现出最适当的结果，而不是某些不道德的网络员制造的、塞满关键字 的无用网页。为了达到此目的，搜索引擎优化应运而生。

**7.4.2** **网页级别**

拉里·佩奇 (Larry Page) 和谢尔盖·布林 (Sergey Brin) 提出了网页评估的新概念， 即利用网页级别来评估网页，网页级别依赖导入链接，并利用导入某网页的链接相当于给 该网页价值投一票的方法来创建系统。导入链接越多就表明该网页越有价值。而导入链接 本身价值直接根据该链接来源的网页级别确定。大多数搜索引擎不公开评估的算法。搜索 引擎可以使用多种因素排列目录，每个因素本身和因素所占比重不断地改变，所以搜索引 擎优化可以使用不同优化方法。

**7.4.3** **搜索引擎优化的方法**

网站管理员为了使自己的网站拥有较好的搜索排名，甚至将不相关的关键字塞在网页 中，用来“欺骗”搜索引擎进而获得好的搜索排名。早期的搜索引擎，如 Altavista 与 Infoseek 就开始调整搜索算法以避免受到人为的干扰，搜索引擎需要不断优化。新站点不需要提交 到搜索引擎，就能登记上市。 一个创建好的、其他站点的简单链接就可使搜索引擎访问新 站点，开始搜索该网站内容， 一旦搜索引擎发现了新站点，它一般将访问和开始索引该站， 直到所有标准的超链接链接到页索引为止。下面介绍几种搜索引擎优化的方法。



**大数据基础与应用**

**1.** **白帽方法**

搜索引擎优化的白帽法主要考虑搜索引擎哪些内容可接受，哪些内容不能接受，即为 用户创建内容，而非搜索引擎，使这些内容易于被蜘蛛机器人索引。网站设计或构建时， 出现致命错误导致排名不好。白帽法企图发现并纠正错误，例如机器无法读取的菜单、无 效链接、临时改变导向或粗劣的导引结构等。因为搜索引擎是以文本为中心，许多有助于 提高网页亲和力的手段同样便于搜索引擎优化。这些方法包括优化图形内容、ALT 属性和 增加文本说明等。

下面是搜索引擎常用的方法：

① 在每页使用一个短的、独特的和相关的标题。

② 编辑网页，用与该页的主题有关的具体术语替换隐晦的字眼，这有助于该站诉求的 观众群在搜索引擎上搜索，而被正确导引至该站。

③ 在站点增加相当数量的原创内容。

④ 使用合理大小、准确描述的汇标，而不过度使用关键字、惊叹号或不相关标题术语。

⑤ 确认所有页可通过正常的链接来访问，而非只能通过Java 、JavaScript 或 Adobe Flash 应用程序访问。这可通过使用一个专属列出该站所有内容的网页实现。

⑥ 通过自然方式开发链接：在混淆不清的指南上，写封电子邮件给网站员，说明刚发 布了一篇很好的文章，并且请求链接，这种做法很可能为搜索引擎所认可。

⑦ 参与其他网站的网络集团，只要其他网站是独立的、分享同样题目和可比较质量 即可。

**2.** **黑帽方法**

搜索引擎管理员认为任何搜索引擎的优化，其目的是用来改进网站的页排名都是垃圾 索引。然而，随着时间流逝，业内公认促进搜索引擎排名与流量结果的某些手段不可接受。 因为搜索引擎以高度自动化的方式运作，网站管理员通常利用某些未被搜索引擎认可的手 段、方法来促进排名。这些方法经常未被注意，除非搜索引擎管理员亲临该站点并注意到 不寻常活动，或在排名算法上的某个小变化导致站点丢失以过去方式取得的高排名。下述 是几个常见的未被搜索引擎认可的手段，如果被发现，该网站有可能将永远从搜索引擎名 单中被剔除。

① 斗篷法是指网站站长用了两版不同的网页来达到优化的效果。 一个版本只给搜索引 擎看， 一个版本给用户看。

② 关键字装进隐秘文本是另外一种欺骗搜索引擎的做法。通常是指设置关键字的颜色 和网页背景颜色一样，或通过隐秘特性来达到优化效果。

③ 桥页也叫门页，通常使用软件自动生成大量包含关键词的网页，然后从这些网页自 动转到主页。目的是希望这些以不同关键词为目标的桥页在搜索引擎中得到好的排名。当 用户点击搜索结果的时候就自动转到主页，也可以在桥页上放一个通往主页的链接，而不 用自动转向主页。

④ 付费链接是利用支付费用方式要求其他网站提供链接至自身网站，借此伪装高信任



第7章 大数据搜索技术

网站来欺骗搜索引擎，付费链接类型多为锚文字类型。

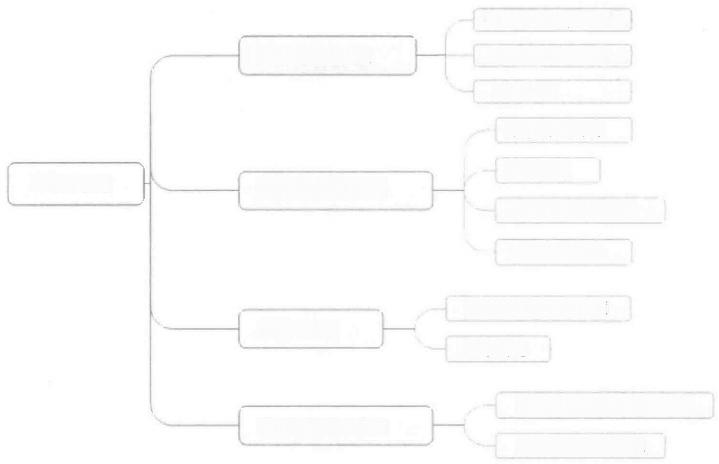
小结

对于大数据技术来说，搜索引擎是一个重要的研究课题，这是由于大数据存于互联网 上，搜索离不开搜索引擎。对于大数据来说，语义搜索引擎、图谱等是重要技术。而搜索 引擎的优化、大数据的流向趋势、搜索引擎的优化等技术都是必不可缺的技术。提高网站 质量，比直接查找、操控搜索排名更具有生命力。语义搜索引擎是未来搜索引擎的发展方 向，语义 Web的发展和自然语言处理技术是其发展的基础。语义搜索引擎设计的最终目标 是使计算机具有人的智能，并以解决问题的形式把结果返回给用户。



**第** **8** **章** **大数据存储**

本章主要内容



大数据存储模型

大数据存储问题

存储方式

数据容量问题

大图数据

大数据的存储技术

分布式存储的架构

数据存储管理

云存储的意义与问题

数据云存储

技术措施

磁盘与磁盘阵列的可靠性

数据存储的可靠性

文件系统的可靠性

大数据存储概述

大数据存储

**8.1** **大数据存储概述**

基于系统软件的角度，大数据存储是对现有的分布式存储的扩展，与单机存储的最大 区别是数据的容量。大数据的容量要扩展到与互联网数据规模(基于互联网产生的数据规 模)相当的级别，但在科学计算大数据中，数据量更大。

数据存储是指数据流在加工处理过程中产生的临时文件或加工处理过程中需要查找的 信息的存储。数据以某种格式记录在计算机内部或外部存储介质上。数据存储需要命名， 命名要反映信息特征的组成含义。数据流反映了系统中流动的数据，表现出动态数据的特 征。需要查找的信息的数据存储反映系统中静止的数据，表现出静态数据的特征。本章介 绍目前常用的几种磁盘存储方式与大数据存储模式。

对于大数据来说，大数据存储和大数据分析两者的关系密切相关。大数据存储是大数 据分析的基础，大数据分析是大数据存储的数据的价值升华。然而，到目前为止，大数据 分析与大数据存储还是计算机的两个不同技术领域。大数据存储关注并研究可以扩展至PB 甚至 EB 数量级别的数据存储平台，而大数据分析关注的是在最短时间内处理大量的不同



第8章 大数据存储

类型的数据集。

**8.1.1** **大数据存储模型**

大数据主要的存储模型分为无格式的文件数据存储和有格式的文件数据存储。

**1.** **无格式的文件数据存储**

无格式的文件数据是指无任何格式的文件数据，也就是说，被存储的文件数据是任意 的二进制流。

**2.** **有格式的文件数据存储**

有格式的文件数据存储是指具有一定格式的数据存储，也就是说，被存储的数据具有 确定的格式，例如关系数据库中的数据就是具有确定的格式或结构，构成二维数据表格的 形式。

**8.1.2** **大数据存储问题**

对于大数据的存储，主要考虑如下问题。

**1.** **容量**

大数据的容量可达到 PB 级的数据规模，因此，对于海量数据存储系统需要有相应等 级的扩展能力。存储系统的扩展要简便，可以通过增加模块或磁盘柜来增加容量，甚至不 需要停机。在解决容量问题上，LSI 公司提出了全新 Nytro 智能化闪存解决方案，可以将数 据库事务处理性能提高30倍，并且每秒超过4.0GB的持续吞吐能力，非常适用于大数据分 析与处理。

**2.** **延迟**

大数据应用存在实时性的问题。特别是涉及与网上交易或者金融类相关的应用，大数 据应用的实时性问题就更显得突出，仅允许小延迟。为了满足延迟指标，很多大数据应用 环境需要较高的 IOPS 性能，例如高性能计算与大规模分布计算等。此外，服务器虚拟化 的普及也导致了对高 IOPS 的需求。针对这些问题，各种模式的固态存储设备应运而生， 小到简单的在服务器内部做高速缓存，大到全固态介质可扩展存储系统通过高性能闪存存 储，促进了自动、智能地读写热点数据的高速缓存技术发展。

**3.** **安全**

某些特殊行业的应用，例如金融数据、医疗信息以及政府重要信息等，都有自己的安 全标准和保密性需求。虽然对于IT 管理者来说这些并没有什么不同，而且都是必须遵从的， 但是，大数据分析往往需要多类数据相互参考，而在过去并不需要这种数据混合访问的情 况。大数据应用催生出一些新的、需要考虑的安全性问题，进而实现了企业级存储的性能 和可靠，既安全又方便。



**大数据基础与应用**

**4.** **成本**

对于正在应用大数据环境的用户来说，成本控制是关键。控制成本表明要使得每一台 设备都实现更高的效率，同时还要减少昂贵的部件。

(1)删除重复数据

删除重复数据能够提高大数据的质量，删除各种数据类型的重复数据技术已经得到广 泛应用，从而提升存储效率，为大数据存储应用带来更多的价值。

(2)减少后端存储的消耗

在数据量不断增长的环境中，需要减少后端存储的消耗。数据中心使用的传统引导驱 动器不仅故障率高，而且具有较高的维修和更换成本。如果使用数据中心的独立服务器引 导驱动器，则能将可靠性提升多达100倍。它对主机系统是透明的，能为每一个附加服务 器提供唯一的引导镜像，可简化系统管理，提升可靠性，并且节电率高达60%,真正做到 了节省成本。

**5.** **长期保存**

大数据应用通常要求数据要保存多年，例如医疗信息通常是为了保证患者的生命安全 需要保存几年或者几十年，而财务信息通常要保存7年。有些使用大数据存储的用户却希 望数据能够保存更长的时间，因为任何数据都是历史记录的一部分，大数据的分析大都是 基于时间段进行的。要实现长期的数据保存，就要求能够持续进行数据一致性检测以及保 证长期高可用的特性，同时还要实现数据直接在原位更新的功能。

**6.** **灵活性**

大数据存储系统的基础设施规模大，因此必须经过仔细设计，才能保证存储系统的灵 活性，并能够使其随着应用分析软件一起扩展。在大数据存储环境中，已经没有必要再做 数据迁移，这是由于数据同时保存在多个部署站点中。 一个大型的数据存储基础设施一旦 开始投入使用，就很难再进行调整，因此它必须能够适应各种不同的应用类型和数据场景。

**7.** **应用感知**

应用感知存储可以根据性能、可用性、可恢复性、法规要求及其价值来调整存储，以 适应存储对应的单个应用。应用感知存储可以优化数据布局、数据行为和服务质量水平， 以确保最佳性能。

应用感知磁盘按照关键任务次序识别并存储数据。例如，公司日常销售库存系统应用 数据都存储在磁盘外部边缘，以确保快速访问。CRM 数据保存到性能中等的磁盘中部，而 电子邮件归档可以存储在磁盘高容量的中心部分。应用感知可以显著提高存储密集型应用 任务的性能，例如归档、备份、灾难恢复和搜索等。在某些情况下，整合和分层存储可以 提高磁盘使用率高达80%～90%。服务质量水平保持不下降的前提下，自动化关键任务的 存储管理时间也减少了，最高可减少50%。

在主流存储系统领域，应用感知技术的使用越来越普遍，它也是改善系统效率和性能 的重要手段，所以应用感知技术也应用在大数据存储环境里。





第8章 大数据存储

**8.** **小用户**

作为商业需求，不仅是大型用户群体应用大数据，小型企业也一定会应用大数据。目 前已经开发出一些小型的大数据存储系统，主要吸引那些对成本比较敏感的用户。

**8.1.3** **存储方式**

**1.** **存储介质**

常用的存储介质为磁盘和磁带。数据存储组织方式因存储介质不同而异。在磁带上， 数据仅按顺序文件方式存取；在磁盘上，数据则可按使用要求直接存取。数据存储方式与 数据文件组织密切相关，其关键在于建立记录的逻辑与物理顺序间对应的关系，确定存储 地址，以提高数据存取速度。

**2.** **直接连接存储**

DAS 是 Direct Attached Storage 的英文缩写，即直接连接存储，是指将外置存储设备通 过连接电缆，直接连接到一台主机上，再连接到存储系统中，使得数据存储是整个主机结 构的一部分。在这种情况下，文件和数据的管理依赖于本机操作系统。操作系统对磁盘数 据的读写与维护管理，需要占用主机资源，如CPU、系 统I/O 等。直接连接存储的优点是中间 环节少，磁盘读写带宽的利用率高，成本也比较低。缺点是其扩展能力有限，数据存储占 用主机资源，使得主机的性能受到相当大的影响，同时主机系统的软硬件故障将直接影响 对存储数据的访问。直接连接存储方式适用于小型网络及一些硬盘播出系统。

(1)直接连接存储的种类与限制

主机与存储设备的连接有多种方式，例如 ATA 、SATA 、SCSI 、FC(Fibre Channel) 等，在实际应用中常采用SCSI 方式， SCSI 所提供的存储服务的主要限制如下。

① 与服务器连接距离有限，不到10米；

② 可连接的服务器数量有限， 一般只有两台，无法服务更大规模和更复杂的应用环境； ③ SCSI 盘阵受固化的控制器限制，无法进行在线扩容。

(2)直接连接存储的环境

与 PC 机存储架构的存储方式一样，将外部存储设备直接挂接在服务器内部总线上， 数据存储设备成为整个服务器结构的一部分。直接连接存储方式主要适用以下环境。

① 小型网络。因为小型网络的规模较小，数据存储量小，而且也不是很复杂，采用这 种直接连接存储方式对服务器的影响不大，并且这种存储方式也十分经济，适合拥有小型网 络的企业用户。

② 地理位置分散的网络。虽然网络规模较大，但在地理分布上分散，通过存储域网络 (Storage Area Network,SAN) 存储或网络连接存储 (Network Attached Storage,NAS) 在 它们之间进行互联非常困难，此时可以将各分支机构的服务器采用直接连接存储方式，这 样可以进一步降低成本。

③ 特殊应用服务器。在一些特殊应用的服务器上，如集群服务器或某些数据库使用的 原始分区，均要求存储设备直接连接到应用服务器，可以采用直接连接存储方式。



**大数据基础与应用**

直接连接存储结构下的数据保护流程相对复杂，如果做网络备份，那么每台服务器都 必须单独进行备份，而且所有的数据流都要通过网络传输。如果不做网络备份，那么就要 为每台服务器都配一套备份软件和磁带设备，备份流程的复杂程度大大增加。

为了应用高可用性的直接连接存储方案，就需要降低解决方案的成本，例如， LSI 的 12Gb/s SAS, 它有直接连接存储，通过直接连接存储能够很好地为大型数据中心提供支持。 对于大型的数据中心、云计算、存储和大数据，都对直接连接存储性能提出了更高的要求， 云中心和企业数据中心数据的爆炸式增长也推动了支持更高速数据访问的高性能存储接口 的需求，因为LSI12Gb/s SAS能够满足这种性能增长的要求，它可以提供更高的IOPS 和更 高的吞吐能力，12Gb/s SAS提高了更高的写入的性能，并且提高了RAID 的整个综合性能。

在服务器与存储的各种连接方式中，直接连接存储是一种低效率的结构，而且也不方 便进行数据保护。直接连接存储无法共享，因此经常使得某台服务器的存储空间不足，而 其他服务器却有大量的存储空间处于闲置状态，无法利用。如果存储不能共享，也不能够 实现容量分配与使用需求之间的平衡。

**3.** **网络连接存储**

网络连接存储 (NAS) 全面改进了低效的 DAS 存储。它采用独立于服务器，单独为网 络数据存储而开发的一种文件服务器来连接存储设备，形成一个网络。这样数据存储就 不再是服务器的附属品，而是作为独立网络节点而存在于网络之中，可由所有的网络用 户共享。

由于 NAS 可无须网络文件服务器，不依赖通用的操作系统，而是采用一个专门用于数据存 储的简化操作系统，内置了网络通信协议，其内嵌的操作系统及硬件体系结构专门针对文 件管理和存储管理进行设计和优化，去掉了通用服务器的大多数计算及多媒体功能，能提 供高效率的文档服务。不仅响应速度快，而且数据传输速率也很高。NAS 按照 TCP/IP 协 议进行通信，以文档的方式进行数据传输，采用业界标准文件共享协议，如NFS 、HTTP 、 CIFS 实现共享。由于NAS 自带文件管理系统，各节点安装不同操作系统，如 APPLE 系统、 Windows系统、Linux 、UNIX 的客户机使用同一文件管理系统，完全实现异构平台之间的 数据级共享。

网络连接存储数据存储是基于现有的企业以太网网络设计，按照 TCP/IP 协议进行通 信，以文件的I/O 方式进行数据传输。

网络连接存储是一种在以太网上实现数据存储的技术， NAS 实际上是一个嵌有网络通 信及文件管理功能的专用存储服务器，具有以下优点。

(1)即插即用

网络连接存储是独立的存储节点存在于网络之中，与用户的操作系统平台无关，真正 实现即插即用。

(2)存储部署简单

网络连接存储不依赖通用的操作系统，而是采用一个面向用户设计的、专门用于数据 存储的简化操作系统，内置与网络连接所需要的协议。因此，整个系统的管理和设置较为 简单。





第8章 大数据存储

(3)存储设备位置非常灵活

网络连接存储是部件级的存储，可直接通过双绞网线连接在 IP网络上，作为网络的一 个节点而存在。它是真正即插即用的产品，并且物理位置灵活，可放置在工作组内，也可 放在其他地点与网络连接。

(4)管理容易且成本低

NAS 具有安装简单、扩展容易、维护方便、成本低等特点。

NAS 系统常用于处理非结构化的数据(比如文档和图像),不适合用于满足事务型数 据的存储需求。由于NAS 采用的是较高端应用层面的 NFS (网络文件系统)协议和CIFS (通用网络文件共享)协议，无形中延长了系统响应的时间，所以，它的数据传输速度较 SAN 要慢一些。其缺点是存储性能较低、可靠度不高。

**4.** **存储域网络存储**

1991年， IBM 公司在S/390服务器中推出了ESCON(Enterprise System Connection) 技术。它是基于光纤介质、最大传输速率达17MB/s 的服务器访问存储器的一种连接方式。 在此基础上，进一步推出了功能更强的 ESCON Director(FC SWitch), 构建了一套最原始 的存储域网络 (SAN) 系统。

存储域网络存储是指通过支持SAN 协议的光纤信道交换机，将主机和存储系统联系起 来，组成一个 LUN Based 的网络，其核心技术就是 Fibre Channel(FC, 光纤信道)协议， 支持HIPPI 、IPI 、SCSI 、IP 、ATM等多种高级协议。FC 是为了解决传统 SCSI 的传输距离 限制而发展起来的一种技术。光纤信道协议的最大特性是将网络和设备的通信协议与传输 物理介质隔离开，这样多种协议可在同一个物理连接上传送。与传统技术相比， SAN 技术 的最大特点是将存储设备从传统的以太网中隔离出来，成为独立的存储局域网络。SAN 使 得存储与服务器分开成为现实。SAN 技术的另一大特点是完全采用光纤连接，从而保证了 大的数据传输带宽，目前其数据传输速度已达4Gb/S, 传输距离可达100km。一条单一的 FC环路最大可以承载126个设备。SAN 具有以下优点：专为传输而设计的光纤信道协议， 使其的传输速率和传输效率都非常高，特别适合于大数据量高带宽的传输要求。由于SAN 采用了网络结构，使其具有无限的扩展能力。SAN 的缺点是成本高，管理难度大。

存储域网络存储方式创造了存储的网络化。存储网络化顺应了计算机服务器体系结构 网络化的趋势。

(1)存储域网络的组成

存储域网络的硬件基础设施是光纤通道，用光纤通道构建的存储域网络由以下三个部 分组成。

① 存储和备份设备：磁带、磁盘和光盘库等。

② 光纤通道网络连接部件：主机总线适配卡、驱动程序、光缆、集线器、交换机、光 纤通道和SCSI 间的桥接器等。

③ 应用和管理软件：备份软件、存储资源管理软件和存储设备管理软件。

(2)存储域网络的优点

① 网络部署容易。





**大数据基础与应用**

② 高速存储性能。因为存储域网络采用了光纤通道技术，所以它具有更大的存储带 宽，存储性能明显提高。存储域网络的光纤通道使用全双工串行通信原理传输数据，传 输速率高。

③ 良好的扩展能力。由于存储域网络采用了网络结构，扩展能力更强。光纤接口提供 了更长的连接距离，实现物理上的分离，即不在本地机房存储数据。

**5.IP-SAN**

(1)ISCSI 的产生

IP-SAN 实际上就是 ISCSI, 由于以下两点原因，导致了ISCSI 的产生。 ① 光纤信道。SAN 不仅昂贵，而且管理复杂。

② 有极少部分的软件由于种种原因不支持网络文件系统，即不能够在 NAS 上无缝使 用。ISCSI(Internet SCSI) 是 IETF (互联网工程任务小组)制定并于2003年2月正式发 布的标准协议，可以理解成SCSI overTCP/IP,即网络上的SCSI。ISCSI协议定义了在TCP/IP 网络发送、接收数据块级的存储数据的规则和方法。发送端将 SCSI 命令和数据封装到 TCP/IP 包中再通过网络转发。接收端收到TCP/IP 包之后，将其还原为 SCSI 命令和数据并 执行，完成之后将返回的SCSI 命令和数据封装到TCP/IP 包中，再传送回发送端。而整个 过程在用户看来，使用远程的存储设备就像访问本地的SCSI 设备一样简单。ISCSI 集合了 SCSI 、以太网和 TCP/IP 等技术，存在很多优点。

(2)ISCSI 的优点

① 基于TCP/IP 协议，基础是传统的以太网和因特网，技术熟悉，维护方便。

② 支持ISCSI 技术的服务器和存储设备能够直接连接到现有的IP 交换机和路由器上， 因此 ISCSI 技术具有易于安装、不受地理限制、良好的互操作性、管理方便等优势。

③ 速度与以太网络速度正相关，IP 网络的带宽发展相当迅速，随着10Gb 以太网络的 普及，其传输速率将会显著提高。

④ ISCSI 实现在 IP 网络上应用 SCSI 的功能，比价格高昂的光纤信道的性价比更高。 虽然ISCSI 前途光明，但目前在性能上仍无法同光纤信道相比，速度仍不理想，而且IP 网 络的传输效率和延迟都不适应实时性要求高的存储数据传输。其次，其受限于 I/O 端口的 速度：在 Host 主机及Target 存储设备两处的I/O 端的速度一直提不上来，所以即使10Gb 以太 网络普及，I/O 端的速度瓶颈仍然会拖累其传输效能。

**6.DAS 、NAS 和** **SAN 三种存储比较**

DAS 、NAS 和 SAN 三种存储共存与互补，可以满足信息化应用。

(1)连接方式

从连接方式上比较， DAS 的存储设备直接连接应用服务器，具有一定的灵活性和限制 性。NAS 通过网络技术 (TCP/IP 、ATM 、FDDI) 连接存储设备和应用服务器，存储设备 位置灵活，随着万兆网的出现，传输速率有了很大的提高。SAN 则是通过光纤通道技术连 接存储设备和应用服务器，具有很好的传输速率和扩展性能。三种存储方式各有优势，相 互共存，占到了现在磁盘存储市场的70%以上。





第 8 章 大数据存储

(2)产品的价格

SAN 和 NAS 产品的价格仍然远远高于 DAS, 许多用户出于价格因素考虑选择了低效 率的直连存储而不是高效率的共享存储。

(3)自动精简配置

SAN 和 NAS 系统可以利用自动精简配置技术来弥补早期存储分配不灵活的问题。与 直连存储架构相比，共享式的存储架构，如 SAN 或者 NAS 都可以较好地完成存储问题。 于是淘汰直接连接存储的进程越来越快。但是，目前直接连接存储仍然是服务器与存储连 接的一种常用的方式。

**8.2** **大数据的存储技术**

随着大数据应用的飞快发展，现已出现了独特的架构，而且也直接推动了存储、网络 以及计算技术的发展。由于大数据处理的需求是一个新的挑战。硬件的发展最终还是需要 软件需求推动，所以大数据分析应用需求正在影响和促进数据存储基础设施的发展。随着 结构化数据和非结构化数据量的持续增长，以及被分析数据的来源多样化，现有的存储系 统已经无法满足大数据存储的需要。基于存储基础设施研究的考虑，现已开始修改基于块 和文件的存储系统的架构设计，以便适应这些新的要求。

**8.2.1** **数据容量问题**

大数据容量为 PB 级的数据规模，因此，海量数据存储系统需要具有一定相应等级的 扩展能力。与此同时，存储系统的扩展一定要简便，可以通过增加模块或增加磁盘柜的方 式来增加容量，甚至不需要停机。基于这样的需求，客户现在越来越多地选择规模可扩展 架构的存储。规模可扩展集群结构的特点是，每个节点除了具有一定的存储容量之外，内 部还具备数据处理能力以及互联设备。与传统存储系统的架构完全不同，规模可扩展架构 能够实现无缝平滑的扩展，避免存储孤岛的出现。

在文件系统中，文件是文件系统的存储单位，大数据除了数据容量巨大之外，文件数 量也十分庞大。因此管理文件系统层累积的元数据是一个困难问题，如果处理不当，将影 响到系统的扩展能力和性能，例如传统的NAS 存储系统就存在这一问题。但是，基于对象 的存储架构就不存在这个问题，它可以在一个系统中管理十亿级别的文件数量，而且还不 能够像传统存储一样遇到元数据管理的困扰。基于对象的存储系统还具有广域性扩展能力， 可以在多个不同的地点部署并组成一个跨区域的大型存储基础架构。

大数据的存储是分布式的存储，并呈现出与计算融合的趋势。由于 TB 、PB 级数据量 的急剧膨胀，传统的数据移动方式已经不适用，导致新的融合趋势的存储服务器的出现。 在这样的架构中，数据不再移动，写入以后分散存储，它的计算节点融合在数据旁边的 CPU 中，数据越来越贴近计算节点，形成了以数据为中心的架构。



**大数据基础与应用**

**8.2.2** **大图数据**

近年来，数十亿顶点规模的大图数据大量出现，不断地对数据管理技术提出了新的要 求。万维网目前已经包含了超过500亿个网页以及数量达万亿级别的统一资源定位符，社 交网络 Facebook 存储的好友网络包含了超过8亿个节点和1000 亿条边。语义网领域的链 接数据规模正迅速呈指数方式提升，目前已包含了310 亿个资源描述框架 (Resource Description Framework,RDF) 三元组以及超过5亿个 RDF 链接。在生物信息学领域，全 基因组序列数据分析的关键环节之一是序列拼接。当前的面向短序列拼接的主流方法是基 于德布鲁因图 (de Brujin)的拼接方法。而人类基因组上的de Brujin 图在最坏情况下具有 420个节点。

图存储技术主要研究图数据在磁盘上存储以及分布式环境下的存储布局形式、划分方 法、复制方法等一系列方法，图存储技术是图数据管理的基石。图的存储方式直接决定了 图数据的访问方式、图查询方式与挖掘的效率。

**1.** **大图存储的基本框架**

大图存储的基本框架是分布式存储框架，其原因如下。

(1)大图数据规模大

10亿顶点规模的大图的每个顶点或者边上存储的附加信息，其规模在TB 级别，甚至 达到 PB 级。

(2)利用了基于分布式内存的计算框架

为了实现对整个图进行随机访问而不是顺序访问，图计算必须基于内存展开。由于当 前内存规模在GB(10° 字节)级别，通过分布式存储就可以直接装入内存，从而降低每台 机器上的图的规模，避免频繁进行磁盘交互。

**2.** **图划分技术**

大图的分布式存储的核心技术是图划分技术。为了将图部署到分布式系统中，需要将 图分为若干部分，而后将每一部分分别存入某一机器。具体而言，图划分是一类问题的集 合，这类问题需要将顶点集合划分为若干单元，这些单元的并集构成总体的顶点集，任意 两个单元相交为空。在考虑通过顶点复制减少通信的策略中，需要放松单元不相交的约束。 通常考虑下述集中因素。

图计算是通过边对顶点进行访问与遍历，跨越机器的边数决定图系统的网络通信开销。 通常将两个端点存储在不同机器上的边称作交叉边。对于非本地数据的访问而导致的网络 通信代价非常高。访问本地内存数据的时间通常以纳秒计算，而网络通信时间则通常以毫 秒计算，两者时间相差巨大。所以图划分的方式直接决定了交叉边的数量，从而决定了基 于该划分方式的图计算所需的通信代价。通常，我们期望交叉边数量最小化，进而降低通 信代价。图划分主要考虑负载均衡与存储冗余问题。

**第8章** **大数据存储**

(1)负载均衡

避免网络通信代价的极端方法是将完整的图信息仅存储于一台机器上。显然，这一方 式很可能超出单台机器的存储上限，同时这一方法也没有并行计算能力。期望图划分的各 个部分具有相近的规模，从而避免负载失衡的情况。负载均衡是相对于机器存储容量和计 算能力而言的。在复杂的实际应用中，可以构建复杂的度量模型用以刻画兼顾机器存储容 量和计算能力的负载均衡模型。

(2)存储冗余

避免网络通信代价升高的另一极端方法是将图的信息在每台机器上复制一份。但这种 方法也容易超出单台机器的存储能力，同时导致大量冗余。对于k 台机器的分布式系统， 这种方法导致 k-1份存储冗余。为了降低冗余，可以选择特定顶点及其邻接信息进行复制。 通常选择度数较大的顶点进行复制，从而在降低通信代价的同时，避免较大冗余。此外， 复制顶点个数和位置的选择等都对最终结果有着直接影响。另外，多个副本之间的一致性 也是重要的问题，通常需要额外的计算代价保持多份副本与主本的完全一致性。

**3.** **大图数据的查询**

如果用图表示社交网络，用户可以看作图的顶点，用户之间的关系(如朋友关系等) 可以看作图的边。与社交网络相类似，Web 网络中的网页可以看作图的顶点，网页之间的 链接关系可以看作图的边。图在社交网络中有着重要的应用。

计算机查询方式经历了文件系统查询、数据库系统查询、 Web 网络查询、社会网络查 询的发展历程，如图8-1所示。



文件系统 数据库系统 Web网络 社会网络

图8-1 查询的演变过程

(1)文件系统

从20世纪60年代开始，计算机开始装配了具有现代意义的操作系统，而文件系统是 操作系统提供的一种存储和组织计算机文件的方法。它提供简单的查询功能，用户可以搜 索文件。

(2)数据库系统

20世纪60年代中期，数据库系统开始应用。20世纪70年代，关系数据模型成为数据 库管理系统的主流系统。70年代后期出现的结构化查询语言 SQL, 极大地提高了数据查询 的灵活性。用户可以通过SQL 语言来进行各种复杂的查询。

(3)Web 网络

从20世纪90年代开始，随着万维网的兴起，Web 搜索引擎被广泛应用。它们通过提 供关键词搜索的功能，使得几乎所有的用户都可以方便地搜索万维网数据。

(4)社会网络

随着 Web 2.0的出现和社会计算的兴起，社交网络系统开始大量应用，如 Facebook、

Linkedln 、 人人网等。图查询技术是适合社会计算的搜索方式。社会计算一般需要考虑社



**大数据基础与应用**

会的结构、组织和活动等社会因素。所有的社会活动构成了社会网络，本质上这是图的一 种表现形式，所以图搜索技术的研究成为关键技术。

**8.2.3** **分布式存储的架构**

大数据带来的变化是从集中走向分布，从以计算为中心转向以数据为中心，分布式存 储的架构是大数据存储架构的发展方向。

**1.** **融合架构**

融合架构方法是面对大数据存储的一个很好的选择。通过融合架构可以实现计算与存 储相融合，进而获得更高的管理效率和更高效能。

从虚拟化、云计算数据保护和融合架构三个维度设计数据中心，可以减少整合的时间 和网络问题判断的时间，并能够实现统一集中透明管理，可以根据工作负载实时动态配置 资源，也可以实时监控，进行故障诊断、排除故障。

融合架构具有不同的形态，其中一种形态是在原来硬件基础上装配一个软件，然后形 成融合架构。实现目的是在线扩展、动态负载平衡，在最大限度提高部署效率前提下，降 低因为硬件问题而导致的应用性能降低和应用的不稳定性。

**2.** **服务级别**

利用移动硬盘就可以存储数据，但是不同于放在数据中心和网络云上的存储享受的服 务级别。为了不使数据成为整个企业发展的负担，而是成为价值增长点，从资料变成资产， 基础架构需要快速、安全地支持新的技术手段，应用级别和服务级别的定义需要有很好存 储架构。集群存储系统是面向实际的应用设计，满足应用分级、资源分层的需求。

**3.PCle 闪存卡**

非结构化的大量数据快速变成信息，不仅要求服务器工作速度快，存储速度也需要与 CPU 的速度相匹配，闪存正是针对当前网络存储速度落后的解决方案，能够有效提高存储 的性能。在云计算、大数据时代，集中式存储管理和维护非常困难，分布式存储模型是大 势所趋。在这其中， PCle 闪存卡、全闪存阵列以及SDK 工具，支持提升各种应用的性能。

SSD 不只是让数据变快，通过SSD 在数据中心的使用，能够帮助节约成本，降低延迟， 加快访问数据的速度，同时还能够提高可靠性和管理级别，结合 DRM 的使用进行软件分 层管理。

大数据分析流程和传统的数据仓库的方式完全不同，其已经变成了业务部门级别和数 据中心级别的关键应用。这也是存储管理员的切入点。随着基础平台变得业务关键化，用 户群更加紧密地依赖这一平台，这也使得其成为企业安全性、数据保护和数据管理策略的 关键研究课题。

用于数据分析平台的分布式计算平台内的存储不是面对的网络连接存储 (NAS) 和存 储区域网络 (SAN), 而是内置的直连存储 (NAS) 以及组成集群的分布式计算节点。这就 使得管理大数据变得更为复杂，因为无法像以前那样对这些数据部署安全、保护和保存流





第8章 大数据存储

程。然而，执行这些流程策略的必要性集成于管理分布式计算集群之中，并且改变了计算 和存储层交互的方式。

**8.2.4** **数据存储管理**

数据无处不在，手机通话记录、商店的 RFID 标签、物流公司的快递产生的数据、银 行的交易数据、出租车的运行轨迹、生活中如此众多的数据记录，凸显现实世界的数据变 化。IDC 的统计表明，到2020年，全球以电子形式存储的数量将达到35ZB(1ZB=100 万 PB), 是2009年存储量的40倍。其中企业数据以55%的速度在逐年增长。

**1.** **传统的数据存储管理不能满足发展要求**

企业的决策者往往根据自己的直觉和经验来规划企业未来的发展战略，而不是依托于 具体的数据。利用数据可以获得有价值的线索，使决策者看到数据分析的重要性。然而从 大数据中抽取数据样本，挖掘数据，形成报表的过程看似简单，但是实际困难较大，涉及 企业IT 系统的各方面，例如企业的数据中心、数据存储、数据管理等多个环节。因此，传 统的数据存储管理已经不能满足大数据的发展要求。

**2.** **大数据存储管理面临的挑战**

电信、金融、零售等行业希望通过大数据的分析手段来帮助做出理性的决策。特别是 电信和金融行业表现更为突出，市场数据没有办法与用户消费数据打通。面临的第一个问 题就是海量数据存储的问题。多数企业正在建设自己的数据中心，来满足大规模的数据量 的产生，但是随着数据的进一步增多，很多数据的查询和分析性能急剧下降，有的数据中 心甚至出现了无法响应的状况，为企业的业务带来了很大损失。

应该考虑数据管理策略对数据进行有效的保护问题，而且在需要时，可使数据随时转 变成价值。只有数据与适合的存储系统相匹配，制定出管理数据的战略，才能低成本、高 可靠性、高效益地应对大量数据。对于企业来说，面临大数据首先解决的问题就是成本和 时间效应问题。为了不错过商机，存储数据管理可以自动删除磁盘和重复数据、备份和归 档，使企业的关键数据存在不同的区域，然后按照特定的业务需求，对数据进行提取、操 作和分析，并形成企业所需要的目标数据。

**3.** **大数据数据管理的意义**

计算机从文字、图像、视频等数据中解析出共性之处，从互联网浩瀚的数据中收获知 识，洞察信息。而数据分析建立在数据管理基础之上。通过交易平台就可以看到整体交易 额的下滑趋势，进而预测到下一年的金融危机的爆发。领军企业与其他企业之间最大的显 著差别在于新数据类型的引入。那些没有引入新的分析技术和新的数据类型的企业，不太 可能成为其行业的领军者。

企业的发展战略与大数据的管理密切相关。信息是企业的财富，如果企业对大数据的 管理适当，利用好大数据，并服务于企业发展战略， 一定能做出明智的决策。现阶段的难 点在于，企业分析的数据仅仅是企业标准化结构数据中的很小一部分，企业未来的数据管





**大数据基础与应用**

理之路还很漫长。

**4.** **大数据的数据管理技术**

在大数据管理技术中，有6种数据管理技术普遍被关注，即分布式存储与计算、内存 数据库技术、列式数据库技术、云数据库、NoSQL 技术、移动数据库技术。其中分布式存 储与计算最受关注，数据管理新技术被关注的程度如图8-2所示。

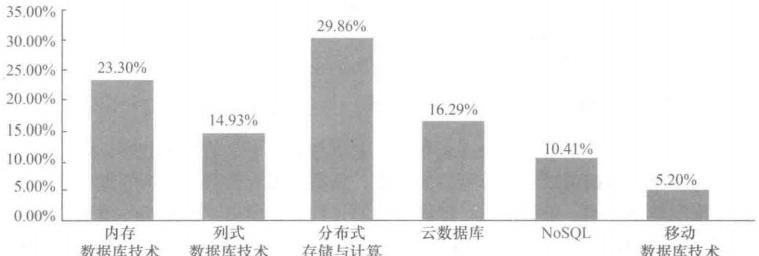


图8-2 数据管理新技术

分布式存储与计算成为最受关注的数据管理新技术，比例达到29.86%,其次是内存数 据库技术，占到23.30%。云数据库排名第三，比例为16.29%。此外，列式数据库技术、 NoSQL 也获得较多关注。从调查结果来看，以 Hadoop 为代表的分布式存储与计算已成为 大数据的关键技术。以 SAPHANA 为代表的内存数据库技术和以 SQLAzure 为代表的云数 据库技术，也将成为占据重要地位的数据管理创新平台。

分布式存储与计算架构可以使大量数据以一种可靠、高效、可伸缩的方式进行处理。 因为以并行的方式工作，所以数据处理速度相对较快，且成本较低，Hadoop 和 NoSQL 都 属于分布式存储技术。

内存数据库技术可以作为单独的数据库使用，还能为应用程序提供即时的响应和高吞 吐量， SAPHANA 是该技术的典型代表。

列式数据库的特点是可以更好地应对海量关系数据中列的查询，占用更少的存储空间， 这也是构建数据仓库的理想架构之一。

云数据库可以不受任何部署环境的局限，随意地进行拓展，进而为客户提供适宜其需求 的虚拟容量，并实现自助式资源调配和自助式使用计量。SQLServer 可以提供类似的服务。

NoSQL 数据库适用于庞大的数据量、极端的查询量和模式演化。企业可以通过 NoSQL 得到高可扩展性、高可用性、低成本、可预见的弹性和架构灵活性的优势。甲骨文在2011 年推出Oracle NoSQL数据库。

移动数据库技术是适应移动计算的产物。随着智能移动终端的普及，对移动数据实时 处理和管理要求的不断提高，移动数据库具有平台的移动性、频繁的断接性、网络条件的 多样性、网络通信的非对称性、系统的高伸缩性和低可靠性以及电源能力的有限性等，受 到业界重视。



第8章 大数据存储

**5.** **大数据的有效管理**

由于数据已经处于核心位置，许多业务已经开始以数据为中心，重新审视业务系统， 希望以此获取大数据带来的价值。但大数据并不是将数据送入仓库就可以了，相反需要更 加精细化的手段管理，才能够做到有效运营数据，具体措施如下。

● 考虑大数据的安全；

● 重新考虑数据解释、分析和预测的能力；

● 建立以数据为导向的数据驱动业务的工作模式；

● 解决流程与数据的矛盾，将流程与数据分离；

● 业务构建以应用为中心转向以数据为中心。

面对不同的数据库和分析环境，企业横向和纵向的扩展能力非常重要。具有简便易行 的横向扩展功能是 Hadoop 迅速应用的原因。其关键在于利用低成本的服务器集群进行大 规模并行处理，比其他的数据管理方式需要更少的专业技能，从而降低了对人员的要求， 能够更经济地实现平滑扩展。

**8.3** **数据云存储**

随着云计算和物联网的迅速发展，越来越多的个人和企业选择将自己的业务迁移到大 型数据中心，以此来降低本地的硬件成本和系统维护费用。在数据的存储从终端转向云端 时，必须考虑云端存储的用户数据规模越来越大，当数据集中在云端之后，必将遇到快速 频繁访问的问题。

**8.3.1** **云存储的意义与问题**

**1.** **云存储的意义**

用户数据集中在云端后，其优点如下所述。

● 用户数据不再隶属于某一个终端；

● 不用担心数据丢失；

● 数据只需存储一份，极大地节省了存储空间。

**2.** **云存储的问题**

云存储需要解决的问题如下所述。

(1)存储成本问题

由于大数据的存储数据规模巨大，即使单位存储空间的成本比传统方式只节省10%, 总体上的经济效益也很可观。

(2)可靠性问题

当数据存储在终端时，数据丢失只影响少数用户，而当数据存储在云端时，数据丢失 将影响大量用户。



**大数据基础与应用**

(3)可用性问题

当数据存储于云端时，如果存储服务出现问题，用户无法访问存储于云端的数据，显 然这对某些应用是不可容忍的。由于云端存储是十分复杂的系统，因此要使云端存储具有 极高的可用性，云端存储的架构实现就必须具有强大的容错能力，能够应对网络、服务器、 IDC 等各种类型的故障，甚至是地震、海啸等自然灾害引起的大范围异常。这种全方位的 抗灾能力一直是云服务面临的困难问题和研究目标。

(4)数据安全问题

用户数据存储于云端时，如果用户敏感数据发生泄露，将带来严重的安全问题，数据 的安全性面临极大的挑战。

(5)访问性能问题

大数据存储于在云端时，需要面对大量用户的同时访问，尤其是多个用户共享一份相 同的数据时，对同一份数据的并发访问明显增加。此外，由于用户在访问云端存储的数据 时必须通过互联网，对互联网带宽提出了更高的要求。

**8.3.2** **技术措施**

**1.** **分布式文件系统**

数据分布存储能够解决数据存储的成本、可靠性、可用性和性能等问题。采用分布式 存储后，可以利用多台廉价的服务器来替代昂贵的专用硬件，降低了成本。将一份数据多 备份，分布存储，可以提高数据可靠性。由于能够同时发挥大量硬盘的 I/O 性能优势，所 以访问性能也将获得极大的提高。

**2.** **重复存储数据**

数据分为结构化数据与非结构化数据。很多数据属于重复的存储数据，要考虑下述 问题。

① 减少系统中无效副本的数量；

② 有效利用数据压缩技术减少空间占用和处理时间；

③ 能够保证数据存储的安全性，针对不同类型的数据，实现高效、便利、有针对性 的存储。

**3.GFS** **与文件系统的区别**

当云端存储涉及成千上万个分布式存储节点时，如何有效地管理这些存储节点并非易 事。由于云端存储几乎不可能容忍停机，如何动态地对系统进行扩容也成为非常复杂的问 题。由于数据被存储多份，数据出现不一致的可能性也大大增加。这些难点都是分布式文 件系统需要解决的问题。谷歌分布式文件系统 GFS 与以往文件系统的区别如下。

● 采用分布式的多台廉价服务器替代专用的存储设备；

● 存储部件故障不再被当作异常，而是将其作为常见的情况加以处理，如存储系统 中的少部分服务器、硬盘或交换机出现故障，都不会影响存储系统的正常运作；



第8章 大数据存储

● 存储的文件可以非常大；

● 大部分文件的更新是通过添加新数据完成的，而尽量不对已存在的数据进行修改 操作；

● 对读多写少的应用进行优化，符合大多数互联网业务的特征。

GFS 满足低成本、高可靠性、高可用性的特征，并且在性能方面可以做到随存储规模 的增加而增大。Hadoop 平台中包含 HDFS 文件系统。HDFS 的出现使得分布式文件系统真 正得到普及。除了 HDFS 外，大型互联网公司也在公司内部开发使用了自主研发的分布式 文件系统，自主研发的主要好处在于可以根据自己的业务特征进行有针对性的优化。

**4.** **存储架构**

在云计算技术出现之前，数据存储的方式很简单，非关系型数据存储在文件系统中， 关系型数据存储到关系数据库中。数据通过各种备份和复制方法保留多份副本。针对不同 的业务场景和需求，使用不同的存储架构来满足需求。为了满足可扩展性需求，针对非结 构化数据，利用分布式文件系统保存海量的数据文件，例如日志信息、话单信息、用户镜 像文件、数据库备份信息和视频文件等。对于文件系统，为了提高性能，要将冷热数据分 离，大小文件分离。根据服务等级协议，还需采用不同的副本及备份策略。 一套完整的文 件系统实际上是由多套子文件系统组合而成。冷热数据分离后，可以针对冷热数据采用不 同的设备和存储策略。热数据基本上保存在内存和SSD 盘内，冷数据使用SATA 硬盘保存， 这样可达到的性价比最高。大小文件分离也有助于选择合适的块大小，减少元数据量。

根据不同的应用情况，结构化的数据存储可细分为联机事务处理OLTP 、OLAP以及最 终一致性的 NoSQL 存储。

**5.NoSQL 技术**

并不是所有的用户数据都适合用文件的形式存储，这也是分布式文件系统不能解决所 有的大规模云端数据存储的重要原因。例如，在微博应用中，需要一次性查询到一个用户 对应的多条博文。如果使用文件存储，每次需要打开多个文件并定位到对应内容的代价将 很高。而如果使用传统的关系型数据库进行存储， 一方面性能不够理想，另一方面又很难 利用分布式系统的优势。因此，NoSQL 系统舍弃了一些 SQL 标准中的功能，取而代之的 是一些简单灵活的功能，使得分布式存储变得更加容易。同时，NoSQL 构建的基本思想就 是尽量简化数据操作，并使操作的执行效率可预估，这也使得分布式的NoSQL 系统有着 非常好的性能。虽然目前分布式的 NoSQL 数据库对分布式支持并不特别完善，但可以通 过应用层逻辑来实现分布式存储。NoSQL 技术目前也还存在一些薄弱环节，主要表现在： 统计分析没有传统关系型数据库简单，NoSQL 数据库的运行与维护的管理工具的成熟度 较差。

**6.** **分布式存储核心算法**

在分布式数据存储中，主要采用的核心算法如下所述。

(1)哈希算法

利用哈希算法可以决定数据存储位置。传统的哈希算法虽然可以实现数据的均匀分布，



**大数据基础与应用**

但是当系统需要扩容时，或者从现有系统中摘除一个节点时，则整个系统的数据需要进行 迁移，消耗的工作量非常大，然而利用一致性哈希算法能够很好地解决这个问题。

(2)容错算法

当数据存储多备份时，如果某个节点无法成功完成读或写， 一种解决方案是把故障节 点摘除。但从实际经验来看，在分布式系统中，网络瞬间故障导致节点无法访问是经常的 事，这种方案引起频繁的节点摘除操作，甚至带来大量的数据不一致。另外一种解决方案 是允许少数节点的读取或写入失败，但在读取时通过综合读取多个节点信息进行校验纠正， 利用了容错技术。

(3)一致性算法

在分布式系统中，如何实现关键状态的一致性是较为困难的环节。例如，有3个控制 节点要自动同步更新一个信息，由于网络故障，其中2个节点同步更新成功了，另外一个 节点更新失败了，这种数据不一致带来了数据错误。 一致性算法就是为了尽可能把发生故 障的可能性降低， Paxos 算法就是一种典型的一致性算法。

**7.** **加密数据云存储模型**

安全与隐私保护成为使用计算的关注问题，云计算存在下述安全风险：特权用户访问、 利用企业免责的法规遵从、数据存储与处理的不确定性、敏感数据的保护、灾难恢复、可 审计和可追踪以及生命周期的有效性。

在加密数据云存储模型中，云存储端除了响应数据共享者的请求并返回加密数据之外， 没有任何密钥生成相关文件。云存储端所有的处理只涉及密文，云存储端并不是可信的第 三方，进而使得实现安全的加密数据云存储和共享变得复杂。但是，如果采用技术与方法 实现可信的第三方加密技术，那么并不能保证第三方不泄露数据。

**8.4** **数据存储的可靠性**

数据的价值不仅在于数据内容本身，更体现在为业务发展所带来的利益、用户体验和 公司盈利状况等。基于数据中心级别考虑，如果存储的数据量十分庞大，则管理系统的复 杂性较高。基于存储设备级别考虑，数据中心为了控制成本，如果大量采用廉价存储设备， 则使得数据极易因硬件设备故障而丢失。因此，如何提高数据中心的数据存储可靠性成了 研究重点。数据存储的可靠性主要包括磁盘阵列的可靠性和文件系统的可靠性两个方面。

**8.4.1** **磁盘与磁盘阵列的可靠性**

作为主流的存储设备的磁盘，其存储容量不断增长，已从最初的 MB 级发展到现在的 TB 级。 一个PB 级的数据中心需要采用上千个磁盘， 一个EB 级的数据中心需要采用上百 万个磁盘。磁盘作为一种机械式存储设备，产生的磁盘错误将导致数据丢失或者损坏。磁 盘错误类型主要包括：磁盘故障、潜在扇区错误和不可检测的磁盘错误。因为磁盘是数据





第8章 大数据存储

的载体，所以磁盘的可靠性与被存储数据的可靠性直接相关。

**1.** **磁盘的可靠性**

磁盘错误分为可检测的错误和不可检测的错误。

(1)可检测的错误

磁盘故障和潜在扇区错误属于可检测的磁盘错误。

① 磁盘故障。磁盘故障指磁盘硬件错误导致整个磁盘数据变得无法访问。磁盘的年故 障率1.7%～8.6%,磁盘的年替换率最高达到13%。由此可以看出，对于由上千个磁盘组成 的 PB 级存储系统来说， 一年内发生故障的磁盘个数就会达到几十个甚至上百个。而对于 由上百万个磁盘组成的 EB 级存储系统， 一年内发生故障的磁盘个数则将达到几万个甚至 几十万个。这是一个惊人的数字。

② 潜在扇区错误。潜在扇区错误指磁盘盘片表面磁介质损坏或 ECC 错误导致某些扇 区数据变得无法访问或不可修复。在32个月内，153万个磁盘中出现潜在扇区错误的磁盘 比例为3.45%,并且这些出现错误的磁盘中潜在扇区错误个数的平均值高达19.7。

(2)不可检测的错误

不可检测的磁盘错误指磁盘固件或硬件错误导致从磁盘读出的数据与磁盘预存的数据 不一致。根据引发的时间点，又可分为不可检测的读错误和不可检测的写错误。不论是不 可检测的读错误还是写错误，在应用中对读操作表现出来的症状相同，要么是读取了过时 的数据，要么是读取了不正确的数据。过时的数据和不正确的数据的区别是过时的数据虽 然正确，但已过期，而不正确的数据就是指错误的数据。不可检测的读错误是瞬时的错误， 而不可检测的写错误则是持久的错误。因此，比较起来，不可检测的写错误对存储系统的 可靠性的影响更为严重。不可检测的写错误主要分为如下三种。

① 误地址写。磁盘固件中的缺陷将使正确的数据被写到错误的位置，从而导致误地 址写。

② 丢失写。如果磁盘的磁头信号强度不够就不能将写数据送达磁盘盘片，并覆盖盘片 上原先存储的旧数据。但此时，如果磁盘又向上层发出“写操作完成”的报告，就会发生 丢失写。

③ 破写。破写是指如果磁盘在写数据的过程中电源被循环重启，就会导致一部分数据 刚写入磁盘时，写操作就结束了。

**2.** **磁盘阵列的可靠性**

为了防止可检测的错误(磁盘故障和潜在的扇区错误),通常采用冗余技术将磁盘组织 成磁盘阵列，冗余技术主要分为多路镜像技术和纠删码技术。

(1)多路镜像技术

多路镜像技术又称为多副本技术，就是将数据复制为多个副本并分别存储，以实现冗 余备份。例如，将每份数据同时存储到t+1 个磁盘中，以防止t 个磁盘同时发生故障。但是 这种方法需要t 倍的额外存储开销，导致存储成本非常高。RAID11 是该技术的一个典型应 用，采用2路镜像并且能容忍1个磁盘发生故障。





**大数据基础与应用**

(2)纠删码技术

纠删码技术的基本思想是将k 个原始数据元素通过编码计算，得到m 块冗余元素，在 所有的 (k+m) 块元素中，任意不高于 m 块元素出错都可以通过重构算法恢复出原来的k 块原始数据。其过程如图8-3 所示，先将 k 个磁盘的原始数据存储到由n 个磁盘组成的磁 盘阵列中(其中 k<n<2k), 并对磁盘阵列进行条带化。然后用一个纠删码对每个条带中的 数据进行编码，以便在多个磁盘同时发生故障时丢失的数据能被恢复出来。如果采用的纠 删码是最大距离可分 (Maximum Distance Separable,MDS) 码，则可以容忍多达m=n-k 个 磁盘同时发生故障。为了容忍t 个磁盘同时发生故障，纠删码技术只需要tn(tVn<1) 倍的 额外冗余。RAID5 是该技术的一个典型应用，采用1个冗余磁盘并且能容忍1个磁盘发生 故障。与多路镜像技术相比，纠删码技术可以提高磁盘阵列的存储效率，并且只引入少量 的额外能耗开销。现已提出多种纠删码技术，主要有RS 码、奇偶校验码和阵列码。

① RS 码。RS 码能够提供高容错能力的最大距离可分码， 但编码和解码工作量较大。RS 码的原理是利用生成矩阵与数 据列向量的乘积得到信息列向量。在重构的时候，利用未出错 的信息列向量所对应的残余生成矩阵的逆矩阵，再与未出错的 信息列向量相乘，来恢复原始数据。磁盘阵列的条带化是独立 构成纠删码算法的信息集合。

② 奇偶校验码。奇偶校验码是一种基于异或运算的纠删 码。最简单的奇偶校验码是单奇偶校验码。在单奇偶校验码中，

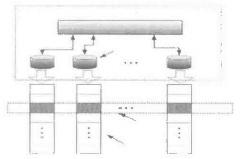


图8-3 磁盘阵列

一个条带中只有唯一的一个校验条块。基于单奇偶校验码可以构造出具有更高容错能力的 奇偶校验码。根据构造方式，奇偶校验码可以分为下述两大类。

● 几何结构奇偶校验码

几何结构奇偶校验码将一个条带中的条块在逻辑上组织成多维几何结构，然后在每个 方向上采用单奇偶校验码进行编码，因此具有规整的结构。虽然此类校验码易于实现，但 是存储效率较低。

● 图结构奇偶校验码

图结构奇偶校验码由基于 Tanner 图构造的低密度奇偶校验码组成。低密度奇偶校验码 最初是为了通信链路的容错而设计，近年来用于广域网的存储。它的优点是容错性好，具 有较高的存储效率，解码复杂度很低。由于它基于不规整的图结构构造，因此在磁盘阵列 中不仅很难实现，还将导致硬件设计复杂化。

③ 阵列码。阵列码是奇偶校验阵列码的简称，其原理是将原始数据和冗余数据都存储 在二维或者多维的阵列中。阵列码具有易于实现、编码和重构的过程相对简单等特点，因此 应用较为广泛。阵列码根据冗余数据和原始数据的排放方式不同可以分为水平和垂直两类。

● 水平阵列码是指冗余数据单独存放在独立的冗余磁盘中，而剩余磁盘存放原始数 据，这种排放方式可扩展性良好。

● 为了使更新复杂度降到最低，提出了垂直阵列码。在 S 垂直阵列码中，某些条块 中既存储了冗余数据，也存储了原始数据。由于垂直阵列码的几何结构简单，因 此其计算开销可均匀地分布到各个磁盘上。



**第** **8** **章** **大数据存储**

● 水平垂直阵列码是一类混合结构的阵列码。

纠删码技术主要用于解决磁盘故障和潜在扇区错误引起的数据丢失问题，而没有考虑 不可检测的磁盘错误引起的无记载数据损坏问题。传统的方法是为磁盘数据附带存储元数 据，包括校验和版本号等额外信息，进而使得无记载数据损坏变得可检测。然而，这种方 法需要额外的存储开销来存取元数据，尤其是在写磁盘数据的时候，还要同步地更新元数 据，从而对写操作性能影响较大。由于增加元数据方法使写操作过程复杂化，增大了写操 作失败的可能性，还可能引起新的无记载数据的损坏。此外，增加元数据方法在检测无记 载数据损坏方面也存在一定局限性，不能检测到所有类型条块的更新错误。

**8.4.2** **文件系统的可靠性**

在大型存储系统中，文件系统是重要的核心子系统。随着磁盘失效的多样性和操作系 统本身的不确定性，文件系统的实现日趋复杂，例如Lustre 文件系统已经达到34万行代码， PVFS2(Parallel Virtual File System Version 2) 文件系统约为21 万行代码。传统的单元测试 需要大量的人力完成，并且几乎无法进行全面的测试，致使分布式文件系统在部署之后， 不可避免地出现错误，进而经常导致系统挂起和系统性能下降，甚至系统崩溃。更为严重 的是，某些软件错误还有可能直接导致数据的丢失或者损坏。为了检查系统某个方面的容 错能力，常采用错误注入方法，检测文件系统是否能应对底层失效，例如磁盘潜在扇区失 效、固件失效等。研究结果表明即使成熟的文件系统，也无法容忍瞬间错误。而且这些错 误还可能被传播到其他的磁盘块，导致系统崩溃。由于一致性检查工具极为复杂，注入的 磁盘错误也有可能导致文件系统无法得到正常修复。

大数据中心的数据存储可靠性取决于磁盘阵列可靠性和文件系统可靠性。提高磁盘 阵列的可靠性方法是应用纠错码技术。提高文件系统的可靠性方法是提供端到端的数据 可靠性。

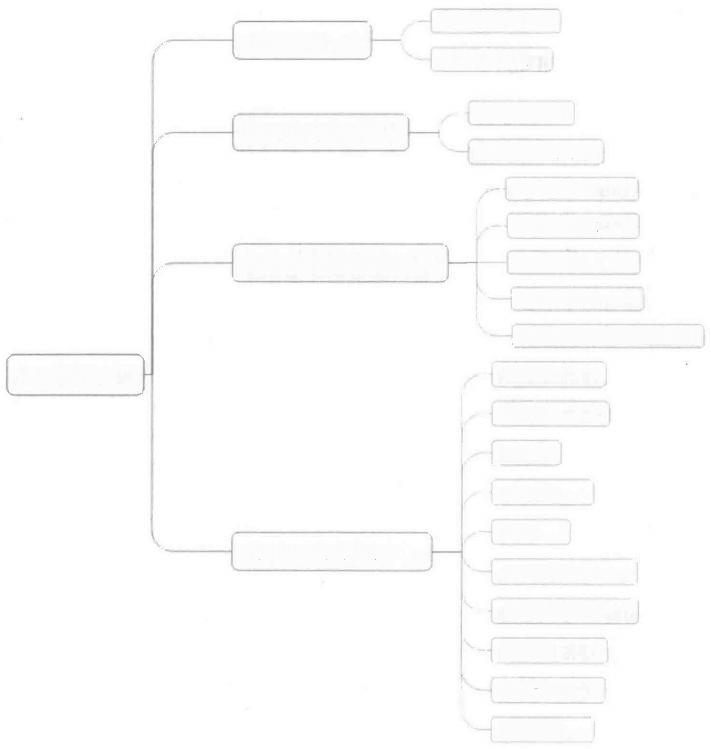
**小结**

本章主要介绍数据存储容量、延迟、安全、成本、数据的积累和灵活性等问题。对集 中常用的存储模型，如存储介质、直接连接存储、网络连接存储、存储域网络存储等做了 较简单的介绍。对大数据存储、大图存储、数据云存储和数据存储的可靠性做了较详细的 介绍。



**第9章** **NoSQL 数据库**

本章主要内容



非结构化问题

NoSQL 概述

NoSQL的产生

NoSQL的特点

NoSQL的特点与问题

NoSQL 面对的问题

键值存储方式

文档存储方式

列存储方式

图形存储方式

存储类型对应的NoSQL 数据库

Cassandra

Lucene

Riak

CouchDB

Neo4j

常用的NoSQL数据库

Oracle的NoSQL

Hadoop的HBase

Bigtable

DynamoDB

MongoDB

NoSQL的主要存储方式

NoSQL数据库

对结构化数据的存储与管理技术已经成熟并得到了广泛应用。由于非结构化是大数据 的重要特征之一，如何将大数据组织成合理的逻辑结构与存储结构是大数据存储与管理中 的一个重要的、必须解决的问题。



第 9 章 NoSQL 数据库

**9.1** **NoSQL 概述**

数据库是按照数据结构来组织、存储和管理数据的仓库，而数据库技术是研究、管理 和应用数据库的技术。NoSQL 代表一族产品及一系列的、不同类型的相互关联的数据存储 与处理技术的集合。

**9.1.1** **非结构化问题**

可计算性问题是指可以使用计算机来解决的实际问题，而计算机本身的优势在于数值 计算，很多非数值问题(例如文字识别、图像处理等)通过转化成为数值问题之后，由计 算机处理。 一个可以使用计算机解决的问题可以定义为可在有限步骤内被解决的问题，哥 德巴赫猜想问题就不属于可计算问题，因为计算机没有办法给出数学意义上的证明，所以 也没有期待计算机能解决世界上所有的问题。分析某个问题的可计算性意义重大，它使得 人们将资源集中在可以解决的问题集中，而不必浪费时间去解决不可能解决的问题，可以 尽早转向使用计算机以外的更加有效的手段去解决。

由于计算机所能处理的问题是离散数值问题，对于连续性非线性复杂问题，需要用有 限元等计算数学工具，把问题映射成计算机能够处理的离散数值计算类问题。随着计算机 功能的日益强大，计算机进入以数据、信息处理为基础的广泛应用领域，最成功的是结构 化数据、信息处理的应用。结构化计算要求每个信息和数据都可以分解为多个有意义的字 段，计算机处理的不是单纯的数值，而是结构数据和信息。

计算机处理能力的增强，也为处理复杂的非结构化信息提供了可能的技术。由于结构 数据与信息所占份额比例小，大多数的数据都是非结构数据，对非结构数据计算与存储的 研究备受重视并已成为研究热点。

**9.1.2** **NoSQL** **的产生**

NoSQL 出现于1998年，是由卡罗·斯特罗兹 (Carlo Strozzi) 开发的一个轻量级的、 开源的、不提供SQL 功能的关系数据库。2009年，约安·奥斯卡森 (Johan Oskarsson) 发 起了一次关于分布式开源数据库的讨论。来自 Rackspace的埃里克·埃文斯(Eric Evans)再 次提出了NoSQL 的概念。这时的NoSQL 主要指非关系型、分布式、不提供ACID 特性的 数据库设计模式。2009 年在亚特兰大举行的 no:sql(east)讨论会是一个新的里程碑。对 NoSQL 最普遍的解释是非关系型的、强调键值存储和文档数据库的优点，而不是单纯的关 系型数据库。

NoSQL 是 Not Only SQL的英文简写，其含义是不仅是结构化查询，所以它是不同于 传统的关系型数据库的数据库管理系统的统称。NoSQL 与 SQL 的显著区别是 NoSQL 不使 用 SQL 作为查询语言。其数据存储不需要固定的表格模式，具有横向可扩展性的特征。 NoSQL 的实现具有两个要点， 一个是主要使用硬盘作为存储载体，另一个是尽可能把随机





**大数据基础与应用**

存储器当作存储载体。

利用关系型数据库中的二维表来存储结构化的数据结构，每个元组字段的组成都相同， 即使每个元组不是都需要所有的字段，但数据库还是为每个元组分配全部的字段。虽然这 种结构便于表与表之间进行连接等操作，但却出现了冗余，这也是关系型数据库性能瓶颈 的一个因素。而非关系型数据库以键值对存储，其结构不固定，每一个元组可以有不一样 的字段，而且每个元组可以根据需要增减自己的键值对，这样就不局限于固定的结构，可 以减少时间和空间的开销，提高存储效率。

随着 Web 2.0 网站的出现与应用，在构建大型的、高并发的社交网络时，应用关系型 数据库出现了下述问题。

**1.** **高并发读写的需求**

Web2.0 网站以及普通的 BBS网站，需要高并发写请求。Web 2.0 网站要根据用户个性 化信息来实时生成动态页面和提供动态信息，因为数据库并发负载非常高，通常可以达到 每秒上万次读写请求，所以基本上无法使用动态页面静态化技术。关系型数据库可以响应 上万次 SQL 查询，但是响应上万次 SQL 写数据请求，硬盘 I/O 无法完成。这是由于对硬盘 的写过程速度比读过程的速度慢。

**2.** **高效率存储和访问的需求**

对于大型的社交网站来说，每天产生大量的用户动态信息。对于关系型数据库来说， 在一张存储2.5 亿条记录的表中进行 SQL 查询，效率极低，甚至不可忍受。此外，大型 Web网站的用户登录系统，面对数以亿计的账号，利用关系型数据库完成困难重重。

**3.** **高可扩展性和高可用性的需求**

在基于Web 的架构中，数据库进行横向扩展困难。当一个应用系统的用户量和访问量 与日俱增时，数据库却没有办法像 Web Server 和 App Server 那样简单地通过添加更多的硬 件和服务节点来扩展性能和负载能力。对于很多需要提供24小时不间断服务的网站来说， 对数据库系统进行升级和扩展困难，通常需要停机维护和数据迁移，可是停机维护随之引 起效率降低。

对于Web2.0 网站来说，关系型数据库的很多主要特性无法发挥。例如很多 Web 实时 系统并不要求严格的数据库事务一致性，对读一致性的要求很低，对写一致性要求也不高。 因此数据库事务管理成了数据库高负载下一个沉重的负担。对关系型数据库来说，插入一 条数据之后立刻查询，可以读出这个数据。但是多数 Web 应用，并不要求这么高的实时性。 对复杂的 SQL 查询，特别是多表关联查询的需求，任何大数据量的 Web 系统都避免多个 大的关联查询，以及复杂报表查询。尤其是社交类型的网站，就需要避免产生这种情况。 更多地采用单表的主键查询，以及单表的简单条件分页查询，极大地弱化了SQL 的功能， 因此，关系型数据库在越来越多的应用场景下已不合适，为了解决这类问题的NoSQL 数据 库应运而生。NoSQL 是非关系型数据库。NoSQL 数据存储不需要固定的表结构，也不存 在连接操作。在大数据存取方面， NoSQL 具有关系型数据库无法比拟的优秀性能。

NoSQL 是非关系型的、分布式的、开源的、横向可扩展的数据库。横向扩展可以将多



第 9 章 NoSQL 数据库

个服务器从逻辑上看成一个实体，在CAP 理论中已有所介绍。NoSQL 主要用于 Web 大规 模的非关系型数据存储。具有模式自由、支持简易复制、简单的 API、最终的一致性、大 容量数据等特性。NoSQL 数据库应用最多的是键值存储方式，还包括文档型存储方式、列 存储方式、图形存储方式、面向对象存储方式与XML 存储方式等。

**9.2** **NoSQL的特点与问题**

CAP 理论、BASE 模型和最终一致性是 NoSQL 数据库存在的三大基石。NoSQL 数据 库满足了数据存储的横向伸缩性的需求。 一些开源的 NoSQL 体系，如 Facebook 的 Cassandra 、Apache 的 HBase 等已得到了广泛应用。

**9.2.1** **NoSQL** **的特点**

NoSQL 的主要特点如下所述。

**1.** **运行在** **PC 服务器集群上**

PC 集群扩充方便并且成本很低，避免了传统数据库共享操作的复杂性和成本。

**2.** **不需要预定义模式**

不需要事先定义数据模式、预定义表结构。数据中的每条记录都可能有不同的属性和 格式。当插入数据时，并不需要预先定义它们的模式。

**3.** **突破了性能瓶颈**

NoSQL 可以省去将 Web 或 Java 应用和数据转换成 SQL 格式的时间，执行速度变得 更快。

**4.** **无共享架构**

相对于将所有数据存储的存储于区域网络中的全共享架构，NoSQL 将数据划分后存储 在各个本地服务器上。因为从本地磁盘读取数据的性能往往优于通过网络传输读取数据， 从而提高了系统的性能。

**5.** **分区不把数据存放于同一个节点**

NoSQL 数据库需要将数据进行分区，将记录分散存储在多个节点上面，并且分区的同 时还要做复制。这样既提高了并行性能，又可以避免单点失效的问题。

**6.** **横向扩展**

随着负载的增加，可将数据库分布到多个不同的主机上。随着每秒事务数与可用性需 求的提高，以及数据库向云或虚拟环境的迁移， NoSQL 数据库从设计之初就考虑了透明横





**大数据基础与应用**

向扩展。可以在系统运行的时候，动态增加或者删除节点。不需要停机维护，数据可以自 动迁移。

**7.** **异步复制**

与RAID 存储系统不同的是， NoSQL 中的复制是基于日志的异步复制。这样，数据就 可以尽快地写入一个节点，而不会被网络传输引起迟延。缺点是并不能总是保证一致性， 这样的方式出现故障时，可能丢失少量的数据。

**8.BASE**

相对于事务严格的ACID 特性，NoSQL数据库保证的是BASE 特性。BASE 是最终一致性。

**9.2.2** **NoSQL** **面对的问题**

NoSQL 得到广泛应用，但还有许多困难需要克服。

**1.** **成熟度**

RDBMS 系统比 NoSQL 系统更加稳定，而且功能也更加丰富，但是 NoSQL 数据库正 在快速发展，许多关键性的功能还有待实现与完善。在应用系统开发中，最经常使用的是 最成熟的技术，而不是最新的技术。

**2.** **支持**

用户希望能得到产品维护的支持与保证，但大多数的NoSQL 系统都是开源项目。虽然 对于每个NoSQL 数据库来说，通常也有一个或多个小型公司对它们提供支持，但是在支持 的范围与资源或可信度方面，不能与Oracle 、Microsoft 或IBM 等大型公司比较。

**3. 分析和商业智能化**

为了满足Web2.0 应用程序的高度可扩展性的要求，NoSQL 数据库的大多数功能都是 面向这类应用程序而设计。但是，在一个应用程序中，具有商业价值的数据已经超出了一 个标准的Web 应用程序需要的“插入一读取一更新一删除”的范畴。在数据库中进行商业 信息的挖掘可以提高企业的效率和竞争力，所以商业智能始终是一个重要的IT 问题。由于 NoSQL 数据库几乎没有提供专用的查询和分析工具，使得一个简单的查询也需要操作者具 有很高超的编程技术，而且常用的商业智能工具无法与 NoSQL 连接。

**4.** **管理**

NoSQL 的设计目标是提供一个零管理的解决方案，目前还远没有达到这个目标。安装 NoSQL 需要很多技巧，维护也需要付出很多的努力。

**5.** **专业知识**

SQL 开发者很熟悉 RDBMS 的概念和 SQL 编程方法，但是，NoSQL 开发者正处于学 习状态。虽然这种情况会随着时间的推移而改变，但是现在找到有经验的 NoSQL 程序员或





第 9 章 NoSQL 数据库

管理员要比找到 RDBMS 专家更为困难。

**9.3** **NoSQL** **的主要存储方式**

在NoSQL 数据库中，最常用的存储方式有文档存储、列存储、键值存储、对象存储、 图形存储和XML 存储等。

**9.3.1** **键值存储方式**

关系型数据库中的表都是存储一些格式化的数据结构，每个元组字段的结构都一样。 即使不是每个元组都需要所有的字段，但数据库会为每个元组分配所有的字段。这样的结 构可以便于表与表之间进行连接等操作，但从另一个角度来说，它出现了冗余，不能满足 以下需求。

● 对数据库高并发读写的需求；

● 对海量数据的高效率存储和访问的需求；

● 对数据库的高可扩展性和高可用性的需求。

为了解决这类问题，非关系型数据库应运而生。Google 的 Bigtable 与Amazon 的 Dynamo 是成功的NoSQL。一些开源的NoSQL 体系，如 Membase 、MongoDB 、Cassandra 、BeansDB、 Redis 等也得到了广泛应用。

**1.** **键值存储的概念**

键值存储方式是NoSQL 数据库中最常用的存储方式，键表示地址，值为被存储的数据， 即一个键值对与地址、地址内容的对应关系如表9-1 所示。

**表9-1** **键值存储方式示意表**

|  |  |
| --- | --- |
| **地** **址** | **被存储的数据** |
| **键** | **值** |

键值存储方式通过对键的一致性查询来获取数据，主要特点如下。

① 存取速度非常快。

② 键值存储适用于不涉及过多数据关系的数据存储，能够有效减少读写磁盘的次数， 比 SQL 数据库存储拥有更快的读写性能，依赖客户端对值进行解释和管理。

③ 键值对可以被创建或删除，与键的值可以被更新。

④ 对不同的编程语言而言，键值存储与哈希表相类似。例如， Java 称之为 HashMap, Perl 称之为 hash,Python 称之为 dict,PHP 称之为 associative array,C++则称之为 boost:unordered\_map<…> 等。

⑤ 键值存储的值中可以存储大量信息。值可以是一个 XML 文档，或者其他任何序列 化形式。





**大数据基础与应用**

**2.** **键值存储方式的分类**

基于键值存储方式建立的数据库的主要特点是具有极高的并发读写性能。例如 Bigtable 、Memcachedb 、Keyspace 等NoSQL 数据库都采用了键值存储方式。键值存储方式 可以分为临时型、永久型和混合型三种形式。

(1)临时型

在临时型的键值存储方式中，将所有的数据都保存在内存中，这样存储和读取的速度 快。但当数据库停止运行时，数据就不存在了，而且，由于数据保存在内存中，所以无法 操作超出内存容量的数据。

(2)永久型

与临时型相反，在永久型的键值存储方式中，数据不会丢失，数据保存在磁盘中。这 就必然对磁盘进行读写操作，所以性能受到影响，但数据不会丢失是其最大的优势。

(3)混合型

混合型的键值存储方式集合了临时型的键值存储方式和永久型的键值存储方式的特 点，进行了折中处理。首先将数据保存在内存中，在满足特定条件，例如默认为15min 一 个以上，或5min 内10个以上，或1min 内10000个以上的键发生变更的时候，将数据写入 到硬盘中。这样既确保了内存中数据的处理速度，又可以通过写入硬盘来保证数据的永久性。 这种存储方式的数据库适合处理数组类型的数据。混合型的键值存储方式总结归纳如下。

① 同时在内存与硬盘上保存数据；

② 可以非常快速地进行保存与读取处理；

③ 保存在硬盘上的数据不会丢失，可以恢复；

④ 适合处理数组类型的数据。

**3.** **举例**

键值数据库的存取效率较高，能够支持大数据中心的数据管理与高速查询，但仅适于 关键字查询，对于非关键字查询和需要逻辑计算的查询，效率低下。

**9.3.2** **文档存储方式**

文档存储支持对结构化数据的访问，但与关系模型不同的是，文档存储没有强制的架构。

文档存储以封包键值对的方式进行存储。在这种情况下，应用对要检索的封包采取一 些约定，或者利用存储引擎的能力将不同的文档划分成不同的集合，以利于管理数据。

**1.** **文档存储方式的主要特征**

(1)文档存储模型支持嵌套结构

例如，文档存储模型支持XML 和 JSON 文档，字段的“值”又可以嵌套存储其他文档。 文档存储模型也支持数组和列值键。这是与关系模型的不同点。

(2)文档存储关注文档的内部结构

这使得存储引擎可以直接支持二级索引，从而允许对任意字段进行高效查询。支持文档





第 9 章 NoSQL 数据库

嵌套存储的能力，使得查询语言具有搜索嵌套对象的能力， XQuery 就是一个例子。MongoDB 通过支持在查询中指定JSON 字段路径实现类似的功能，这是与键值存储的不同点。

(3)存储的内容为文档型数据

不定义表结构，但可以像定义表结构一样使用。关系数据库在变更时比较费事，而且 为了保持一致性还需要修改程序。但是文档存储方式的存储可以省去这些事情，方便、快 捷。文档存储方式与键值存储方式不同的是，文档存储方式可以通过复杂的查询条件来获 取数据。虽然不具备实务处理和 JOIN 这些关系数据库所具有的处理能力，但除此以外的 其他处理基本上都能实现。

(4)主要解决的问题

主要解决的问题不是高并发读写操作，而是实现海量数据存储，同时具有优异的查询 性能。

**2.** **文档存储方式的** **NoSQL** **数据库组成**

BaseX 、CouchDB 、eXist、Jackrabbit、Lotus Notes、MarkLogic Server、Terrastore、MongoDB 和 OrientDB 数据库都是典型的文档存储方式，文档存储方式的NoSQL 数据库主要由文档、 集合、数据库组成。

① 文档相当于关系数据库中的一条记录；

② 多个文档组成一个集合，集合相当于关系数据库的表；

③ 将多个集合在逻辑上组织在一起就是数据库。

**例9-1** 文档数据库 (MongoDB) 中的一个文档为：



1. {

2."name":"zhang",

3."scores":[75,99,87.2]

4. ]

从上述例子中，我们可以看出文档存储方式的下述主要特性：

① 不定义表结构

虽然文档存储NoSQL 数据库不定义表结构，却可以像已定义了表结构一样使用。关系 型数据库在变更表结构时比较麻烦，而且为了保持一致性还需要修改程序。但文档存储的 NoSQL 数据库可以省去这些步骤，确实方便、简单而快捷。

② 单个文档中存在相关数据

NoSQL 数据库并不是绝对没有相对关系结构的记录文件的集合，而是在记录中存储的 数据包含带有节点的树状数据。数据库中的每个记录都以文档形式存在，具备自我描述的 功能，并且独立于任何其他文档。

**例9-2** 文档数据库 (CouchDB) 中的一个文档为：



1.{

2."Subject"."I like Plankton"

3."Author":"Rusty"

4. "PostedDate":"5/23/2006"





大数据基础与应用

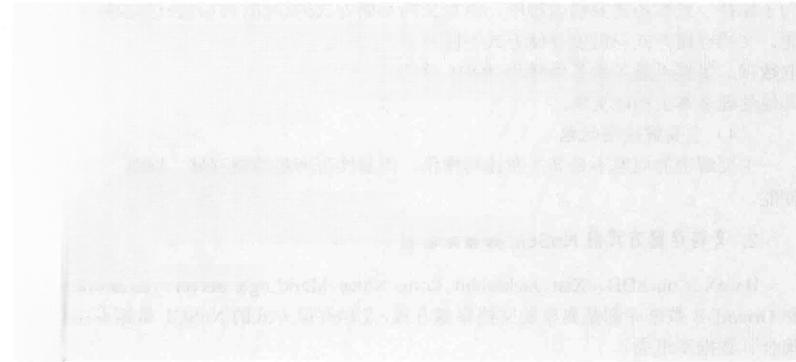


5. "Tags":["plankton","baseball","decisions"]

6."Body"."I decided today that I don't like baseball.I like plankton."

7.]

**例9-3** 文档数据库 (RavenDB) 中的一个文档为：



1.{

2. "BlogPostTitle":"LINQ Queries and RavenDB",

3. "Date":"VDate(1266953391687+0200)V",

4."Content":"Querying RavenDB isvery familiar for.NET developers who are already 5. using LINQ for other purposes",

6. "Comments":[

7. {

8. "CommentorName":"Julie",

9. "Date":"VDate(1266952919510+0200)V",

10."Text":"Thanks for using something l already knowhow to

11.workwith!",

12."Userld":"users/203907"

13.},

14.]

15.}

在例9-3 中，文档中包含字符串数据、数字和数组的简单结构，还可在对象中内嵌入

对象，以获得更复杂的文档结构。

**3.** **唯一键**

主键不能重复，不能为空，唯一键不能重复，可以为空。主键可用于标识数据库中一 条记录，所以不可以为空。唯一键是用来标识数据库中一条记录是否为唯一，所以可以为空。

所有数据库都需要键。如果不提供键，系统则将自动地在内部创建一个键。键对于数 据库的索引功能至关重要。

**9.3.3** **列存储方式**

列存储方式的NoSQL 数据库可以实现按列存储数据，最大的特点是便于存储结构化和 半结构化数据，便于数据压缩，对于某一列或某几列的查询具有非常大的 I/O 优势。例如 Cassandra 、HBase 和 Hypertable 等数据库都是列存储方式的NoSQL 数据库。Bigtable 表面 上与关系型数据库类似，但实际上一个是按行存储数据(关系数据库),另一个是按列存储 数据(列数据库)。因为列数据库中不存在关系，所以不能将适用于关系型数据库的解决方 案运用于列存储方式的数据库。

**1.** **列存储的概念**

列存储方式能够高效率地存储数据，如果列值不存在就不存储，这样一来遇到 null 值 时就能避免浪费空间，减少冗余。在列存储方式中，列族 (Column Family)、超 列 (Super Column) 和列 (Column) 是三个重要的概念。列族和超列在数据库中不占存储空间，也就





**第** **9** **章** **NoSQL** **数据库**

是说如果它们没有值，它们只占用0个字节，列族与传统数据库中的表类似，但和表又不 一样，在列族中的唯一定义是名称和键排序选项。

● 列：由名称、值和时间戳组成，可以忽略时间戳，可当作一个键值对。

● 超列： 一个字典，它是一个包含其他列的列，但不能包含另一个超列。

● 列族： 一个列族中的所有数据都将保存在相同的文件中， 一个列族可以容纳多个

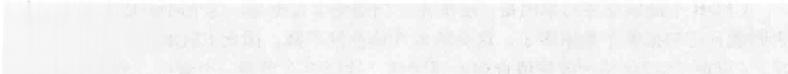
超列或列。

通常将列存储方式的 NoSQL 数据库称为列族数据库 (CFDB) 。 在列族数据库中，如 果列族与超列内部没有值，它们只占用0个字节，列族与传统数据库中的表类似，但与表 又不一样，在列族中的唯一定义是名称和键排序选项，没有任何模式。列族数据库是围绕 真实的物理模型设计，消除了抽象设计。如果架构不正确，那么就无法获取数据，因为列 族数据库是分布式的，所以列族数据库按照键或键范围两种形式进行查询。又因为数据是 基于列族的排序顺序存储的，除了选择升序和降序外，没有办法改变排序方式，键决定了 数据的真实物理位置。与关系数据库中的排序顺序不一样，列数据库的排序顺序不受列的 值影响，只受列名的影响。

**2.** **举例**

**例9-4** 每个数据单元可以作为一组键值对集合，集合本身通过主键标识，又将主键 称为行键。另外，单元按顺序排列进行存储。数据单元按行键排序。例如，下述一张表存 储与人相关的信息。其中包含以下各列：名字(name)、 职业(occupation)、邮 编(post\_code)、 性别 (gender)。

表中某人信息如下：



name:liuming

post\_code:100083

gender:male

同一张表中的另一组数据：



name:liyong

post\_code:102200

第一个数据的行键为1,第二个数据的行键为2。列存储方式的NoSQL 数据库将行键

为1的数据存储在行键为2的数据前面，而且两条数据相互比邻。此外，每条数据中只有 合法的键值对才会被存储。数据按列族存储。列族通常是在配置或启动时定义，列不需要 预先定义或声明，列可以存储任何数据类型，前提是数据能够持久化成 byte 数组。假设上 述例子中列族name 的成员包括列Name, 列 族location 的成员包括 post\_code, 列族 profile 的成员包括 gender。这样一来，上例的底层存储由3个存储桶组成：Name 、location 和profile。 每个桶只会保护合法的键值对，因此，列族 Name 桶中存储下列值：



For row-key:1

name:liuming

For row-key:2





大数据基础与应用

name:liyong

列族 location 桶中存储下列值：



For row-key:1

post\_code:100083

For row-key:2

post\_code:102200

列族 profile 桶只有行键1对应的数值，则它存储为：



For row-key:1

gender:male

在实际存储中，物理上一条数据的列族并不相互隔离，同一行键的所有数据存储在一

起。列族可以代表成员列的键，而行键则代表整条数据的键。

数据按顺序连续存储，当数据逐渐填满一个节点后，它会拆分成多个节点。数据不仅 在每个节点上是有序，而且还跨越多节点形成一个更大的有序集。数据持久化方面会有容 错的考虑，每份数据同时维护3个副本。利用分布式文件系统将数据持久化存储到磁盘上， 分布式文件系统支持将数据存储到集群的多台服务器上。因为有序，数据按行键查找效率 极高，数据访问随机性更小，查找也更简单，就是在序列中查找包含数据的节点。数据插 入发生在队列的尾部，数据更新则在原地进行，不过一般是添加数据的一个新版本到指定 单元中，每个单元始终维护多个版本，版本属性通常可配置。之所以有这么多的限制，主 要原因是 CFDB 设计目标就是运行在多台机器上，存储大量信息，在关系型数据库中几乎是 不可能存储这么多数据的，即使是像 Oracle RAC 这样的集群数据库也存储不了那么多数据。

CFDB 不提供连接的原因是，连接需要扫描整个数据集，这个时候要么使用一个视图， 否则就只有扫描整个数据库了。这会造成性能急剧下降，因此 CFDB 必须避免出现这种情 况。CFDB 不提供按列或按值查询的原因是，这样要么需要一个索引，要么扫描整个数据 集 ，CFDB 将查询限制到只能按键进行查询，确保它能准确地知道查询应在哪个节点上运 行，这意味着每个查询只会扫描一小段数据集，性能自然会好很多。

HBase 是 Bigtable 的开源山寨版本，是建立在 HDFS 之上，提供高可靠性、高性能、 列存储、可伸缩、实时读写的数据库系统。它介于NoSQL 和 RDBMS 之间，仅能通过主键 和主键的范围来检索数据，仅支持单行事务(可通过 hive 支持来实现多表join 等复杂操 作)。主要用来存储非结构化和半结构化的松散数据。与 Hadoop一 样 ，HBase 目标主要依 靠横向扩展，通过不断增加廉价的商用服务器，来增加计算和存储能力。HBase 中的表一 般有这样的特点：

● 一个表可以有上亿行，上百万列。

● 面向列：面向列(族)的存储和权限控制，列(族)独立检索。

● 稀疏：对于为空(null)的列，并不占用存储空间，因此，表可以设计得非常稀疏。

例9-5 Bigtable 采用了键值存储方式，是非关系的分布式数据存储系统。Bigtable 的 设计目的是为了可靠地处理 PB 级别的大数据，并且能够部署到大量计算机上。

对于 Bigtable 而言，数据是没有格式的，就是数据没有表，由用户自己去定义表。用





**第** **9** **章** **NoSQL 数据库**

户也可以自己推测底层存储数据的位置相关性，例如树状结构，具有相同前缀的数据存放 位置接近。在读取的时候，可以把这些数据一次读取出来。数据的下标是行和列的名字， 名字可以是任意的字符串。Bigtable 将存储的数据都视为字符串，但是 Bigtable 本身不去解 析这些字符串，客户程序通常把各种结构化或者半结构化的数据串行化到这些字符串中。 通过仔细选择数据的模式，客户可以控制数据的位置相关性。最后，可以通过 Bigtable 的 模式参数来控制将数据存放于内存中还是硬盘上。

(1)主要特点

① 适合 PB 级数据；

② 分布式、并发数据处理，效率极高；

③ 易于扩展，支持动态伸缩；

④ 适用于廉价设备；

⑤ 适合于读操作，不适合写操作；

⑥ 不适用于传统关系数据库。

(2)应用领域

Bigtable 数据库为搜索、地图、财经、打印、社交网站、视频共享网站和博客网站等业 务提供技术支持。

(3)数据模型

Bigtable 的键值存储方式是稀疏的、分布式的和多维的排序映射。Bigtable 的键具有三 个维数，分别是行键、列键和时间戳，行键和列键都是字节串，时间戳是64位整型数，而 存储的值是一个字节串。 一条键值对记录的格式描述如下：

(row:string,column:string,time:int 64)→string

Bigtable 的行键可以是任意字节串，通常有10～100字节，行的读写都是原子性的。按 照行键的字典序存储数据。Bigtable 的表根据行键自动划分为片，片是负载均衡的单元。最 初表都只有一个片，随着表不断增大，片会自动分裂，片的大小控制在100MB～200MB。 行是表的第一级索引，可以把该行的列、时间和值看成一个整体，简化为一维键值映射。

持久化是指 Bigtable 的数据最终会以文件的形式放到GFS 去。Bigtable 建立在 GFS 之 上本身就意味着分布式，当然分布式的意义并不仅限于此。稀疏是指一个表中不同的行、 列可能完全不一样。



table{

"1":{sth.},//一行

"xxxxx":{sth.}

"yyyyy":{sth.},

"xyzhh":{sth.},

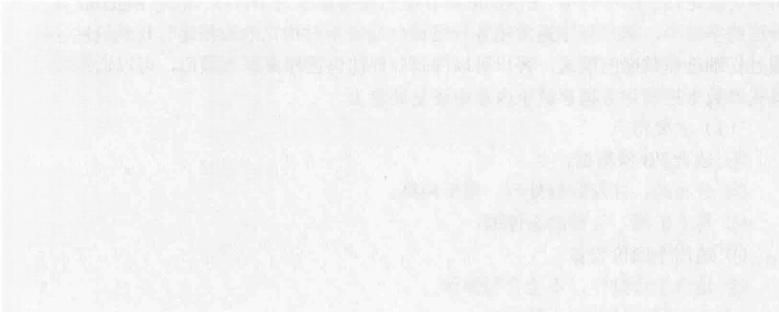
"zzzzz":{sth.}

列是第二级索引，每行拥有的列不受限制，可以随时增加或减少。为了方便管理，列 被分为多个列族 (column family), 是访问控制的单元。 一个列族中的列一般存储相同类型



大数据基础与应用

的数据。 一行的列族很少变化，但是列族中的列可以随意添加删除。列键按照 family:qualifier 格式命名。这次将列拿出来，将时间和值看成一个整体，简化为二维键值映射，类似于：



|  |  |
| --- | --- |
| table{ |  |
| /1 … |  |
| "xxxxx":{ | /1一行 |
| "A:foo":{sth.} | /|一列 |
| "A:bar":{sth.} | /1一列 |
| "B:":{sth.} | /l一列，列族名为 B, 但是列名是空字串 |
| }, |  |
| *"yyyyy":{* | //一行 |
| "A:foo":{sth.}, |  |
| "B:":{sth.} |  |
| }, |  |
| /1 .… |  |
| } |  |

或者可以将列族当作一层新的索引，类似于：

table{

/1 …

"xxxXX":{

"A":{

"foo":{sth.},

"bar":{sth.}

},

"B":{

"":{sth.}

}

*"yyyyy":{*

"A":{

"foo":{sth.},

},

"B":{

"":"ocean"

/1….

/一行

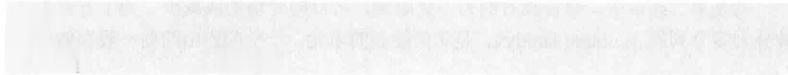
//列族 A

//一列

//列族 B

//一行

时间戳是第三级索引。Bigtable 允许保存数据的多个版本，版本区分的依据就是时间戳。 时间戳可以由 Bigtable 赋值，代表数据进入 Bigtable 的准确时间，也可以由客户端赋值。 数据的不同版本按照时间戳降序存储，因此先读到的是最新版本的数据。加入时间戳后， 就得到了Bigtable 的完整数据模型，类似于：

table{

/…

"xxxXX":{

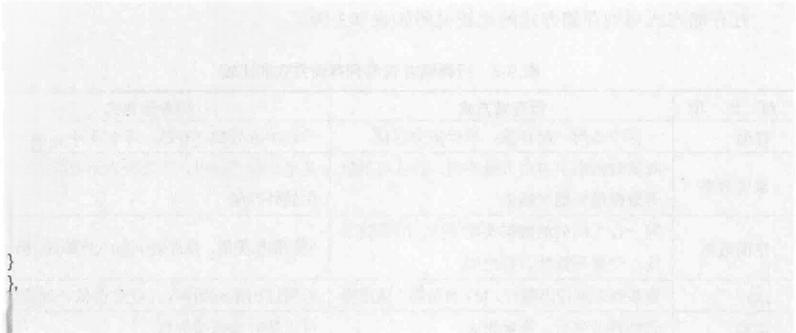
//一行



*数据库*

*第* *9* *章*

*NoSQL*

"A:foo":{

15:"y",

4:"m"

}

"A:bar":{

15:"d"

},

"B:":{

6:"w"

3:“o”

1:"w"

/l

//...

}

//一列

//一个版本

||一列

//一列

查询时，如果只给出行列，那么返回的是最新版本的数据；如果给出了行列时间戳， 那么返回的是时间小于或等于时间戳的数据。例如，查询"xxxxx"/“A:foo”, 返回的值是"y"; 查 询 "xxxxx"/"A:foo"/10, 返回的结果就是"m”; 查询 ”xxxxx"/"A:foo"/2, 返回的结果是空。

**3.** **行存储方式与列存储方式的比较**

传统的关系型数据库主要采用行存储方式，大数据的高效存储与访问要求引发了从行

存储模式向列存储模式的转变，样例数据库如表9 - 2所示。

**表9-2** **样例数据库**

|  |  |  |  |
| --- | --- | --- | --- |
| **用** **户** | **生** **日** | **聊** **天** **记** **录** | **日均在线增长** |
| 用户1 | 1980-5-19 | Xxxxyyyy. | 2 |
| 用户2 | 1988-7-22 | Mm nnnn | 3.9 |

其行存储为：

|  |  |  |  |
| --- | --- | --- | --- |
| 用户1 | 1980-5-19 | Xxxxyyyy… | 2 |
| 用户2 | 1988-7-22 | Mm nnnn. | 3.9 |

其列存储为：

|  |  |
| --- | --- |
| 用 户 1 | 1980-5-19 |
| 用 户 2 | 1988-7-22 |

|  |  |
| --- | --- |
| 用户1 | Xxxxyyyy… |
| 用户2 | Mm nnnn… |

|  |  |
| --- | --- |
| 用户1 | 2 |
| 用户2 | 3.9 |





**大数据基础与应用**

行存储方式与列存储方式的比较说明如表9-3所示。

**表9-3** **行存储方式与列存储方式的比较**

|  |  |  |
| --- | --- | --- |
| **对** **比** **项** | **行存储方式** | **列存储方式** |
| 存储 | 一行中各列一起存放，单行集中存储 | 一行中各列独立存放，单列集中存储 |
| 索引效率 | 海量数据索引占用大量空间，索引效率随 着数据增长越来越低 | 基于列自动索引，大数据查询效率高，不产 生额外存储 |
| 空间效率 | 同一行不同列的数据类型不同，压缩效率 低，空置列依然占据空间 | 列同数据类型，压缩效率高，空置不占据空间 |
| I/O | 查某列必须读出整行，I/O负荷高、速度慢 | 只需读出某列数据，I/O负荷低、速度快 |
| 结构 | 表结构改变后，影响很大 | 可以随时动态增加列 |
| 适用场景 | 数据写入后需要修改和删除，基于行的反 复查询，多用于OLTP数据库 | 批量数据一次写入和基于少量列的反复查 询，多用于OLAP数据库 |

**9.3.4 图形存储方式**

图形数据库存储顶点和边的信息，有的支持添加注释。图形数据库可用于对事物建模， 如社交图谱、真实世界的各种对象。IMDB(Internet Movie Database) 站点的内容就组成了 一幅复杂的图像，演员与电影彼此交织在一起。图形数据库的查询语言一般用于查找图形 中断点的路径，或端点之间路径的属性。

图形存储方式的NoSQL 是图形关系的最佳存储方式，如果使用传统的关系型数据库来 解决图形存储，其性能低下，而且设计困难。例如AllegroGraph 、DEX 、Neo4j 和 FlockDB 等数据库都是图形存储方式的 NoSQL 数据库。

**1.** **属性图信息**

在图的领域，存在着很多不同类型的图模型。按照该模型，属性图中信息的建模使用 下述3种构造单元：

● 节点(即顶点);

● 关系(即边):具有方向和类型(标记和标向);

● 节点和关系上面的属性(即特性)。

更特殊的是，这个模型是一个被标记和标向的属性多重图。被标记的图每条边都有一 个标签，它被用来作为那条边的类型。有向图允许边有一个固定的方向，从末节点(或源 节点)到首节点(或目标节点)。属性图允许每个节点和边有一组可变的属性列表，其中的 属性是关联某个名字的值，简化了图形结构。多重图允许两个节点之间存在多条边。这意 味着两个节点可以由不同边连接多次，即使两条边有相同的尾、头和标记。

**2.Neo4j 图形数据库**

图形结构的数据库同其他行列以及刚性结构的 SQL 数据库不同，它是使用灵活的图形 模型，并且能够扩展到多个服务器上。NoSQL 数据库没有标准的查询语言 (SQL), 因此进





第 9 章 NoSQL 数据库

行数据库查询需要制定数据模型。许多NoSQL数据库都有REST 式的数据接口或者查询API。

常用的图形数据库有Neo4j 、InfoGrid 、Infinite Graph 等。典型应用于社交网络，推荐 系统等，专注于构建关系图谱。

数据模型为图结构，能够更有效利用图结构相关算法，其不足之处是需要对整个图做 计算才能得出结果，不容易做分布式的集群方案。

NoSQL 数据库在以下这几种情况下比较适用。

● 数据模型比较简单；

● 需要灵活性更强的 IT 系统；

● 对数据库性能要求较高；

● 不需要数据的高度一致性；

● 对于给定key, 比较容易映射复杂值的环境。

**9.3.5** **存储类型对应的** **NoSQL** **数据库**

如表9-4所示的内容就是各种典型的存储类型所对应的NoSQL 数据库产品，但它们之 间也有交叉的情况，比如 Tokyo Cabinet/Tyrant 的 Table 既可以划为键值存储类型，又可以 理解为文档存储类型。也可以按使用语言类型不同分类， NoSQL 数据库常使用的语言有 C/C++ 、Java 、Erlang 等，每种NoSQL 数据库都有独到之处，可以根据实际需要，合理地 选择使用。

**表9-4** **存储类型及其所对应的典型** **NoSQL** **数据库**

|  |  |
| --- | --- |
| **存** **储** **类** **型** | **典型的NoSQL数据库产品** |
| 文档存储方式 | MongoDB |
| 列存储方式 | HBase Cassandra Hypertable |
| 键值存储方式 | Redis Tokyo Cabinet/Tokyo Tyrant Flae |
| 对象存储方式 | Db4o Versant |
| 图形存储方式 | Neo4j FlockDB |
| XML存储方式 | Berkeley DB XML BaseX |

工欲善其事，必先利其器。现在已经出现了很多成熟的 NoSQL 产品，表9-4所示 仅为常用的NoSQL 数据库系列，例如 Cassandra 、HBase 、MongoDB 、CouchDB等，表9-5 中也描述了这些 NoSQL 数据库的系列。

数据库主要分为 Bigtable(Google) 和 Dynamo(Amazon) 两种系列。

可伸缩性是分布式系统的一个重要标志。分布式系统设计遵循 CAP 定理，从 EJB 到 MySQL5, 都试图实现CAP 三个核心需求的完美结合，关系型数据库主要满足一致性和可 用性，特别是要求极其严格的一致性，主要是通过数据库锁或 JTA/JDBC 事务实现，过于 苛刻的一致性要求，很难实现分区错性。而NoSQL 数据库，则分为满足一致性和分区错性

(例如 MongoDB 等 NoSQL 数据库)和可用性和分区错性(例如 Dynamo 、Cassandra 等 NoSQL 数据库)两种。





大数据基础与应用

**9.4** **常用的** **NoSQL 数据库**

在大数据技术中，非关系型数据库技术必不可少，但是关系型数据库也是不可或缺的。 关系型数据库与非关系型数据库的比较如表9-5所示。

**表9-5** **关系型数据库与非关系型数据库的比较**

|  |  |  |
| --- | --- | --- |
| **对** **比** **项** | **非关系型数据库** | **关系型数据库** |
| 定义 | (1)无标准定义  (2)主要包括列数据库、键值数据库和文档 数据库等 | (1)创建在关系模型基础之上，借助集合 代数概念与方法来处理数据库中的问题 (2)关系模型由关系数据结构、关系操作 集合、关系完整性约束三部分组成 |
| 接口语言 | (1)无统一标准  (2)包括各自定义的API、类SQL、MR等 | 利用SQL对数据库中的数据进行查询、 操作与管理 |
| 典型案例 | HBase、MongoDB、Redis | Oracle、DB2、Sybase、SQL Server等 |

在数据存储方面要求应用架构具备高伸缩性的横向扩展能力，而NoSQL 正致力于改变 这一现状。关系型数据库给开发者强加了太多东西，它们强行修改对象数据，以满足数据 库系统的需要。在NoSQL 的开发者看来，基于NoSQL 的数据库替代方案正是所需要的。

随着大数据的产生与不断发展，非关系型数据库现在成了一个极其热门的新领域，非 关系型数据库产品的发展非常迅速。现今的计算机体系结构在数据存储方面要有强大的横 向扩展性，而NoSQL 也正是致力于改变这一现状。Google 的 Bigtable和Amazon 的Dynamo 的就是 NoSQL 型数据库，本节简单介绍9种典型 NoSQL 数据库。

**9.4.1** **Cassandra**

Cassandra 最初由 Facebook 开发，后来成了 Apache 开源项目，它是一个网络社交云计 算方面理想的数据库。它集成了其他的流行工具，如 Solr, 现在已经成为一个完全成熟的 大型数据存储工具。Cassandra 是一个混合型的非关系的数据库，类似于 Google 的 Bigtable。 其主要功能比 Dynomite (分布式的 Key-Value 存储系统)更丰富，但支持度却不如文档存 储 MongoDB 。Cassandra 的主要特点就是它不是一个数据库，而是由一堆数据库节点共同 构成的一个分布式网络服务，对 Cassandra 的一个写操作，会被复制到其他节点上，而对 Cassandra 的读操作，也会被路由到某个节点上面去读取。在最近的一次测试中， Netflix 建 立了一个包含288个节点的集群。

**9.4.2** **Lucene**

Lucene 是 Apache 软件基金会开发的一个开放源代码的全文检索引擎工具包。它不是



第 9 章 NoSQL 数据库

一个完整的全文检索引擎，而是一个全文检索引擎的架构。主要使用Lucene 来检索大量的 文本块，它采用了与其他 NoSQL 数据存储相似的模型。如果说查询并不是仅仅局限于精确 的匹配，而是寻找出那些出现在块中的字或者字段的话， Lucene 是最好的查询方式。

**9.4.3** **Riak**

Riak 是一个类似 Dynamo 的分布式键值系统。采用了分布式水平扩展性与高容错性。 可以使用JavaScript 或者 Erlang 来做 MapReduce 查询，它们将查询每个节点，收集结果。 该系统还为类似于Solr 的搜索提供全文索引，同时还提供一个控制面板，可以查看集群的 信息。

**9.4.4** **CouchDB**

CouchDB 不是一个传统的关系型数据库，而是面向文档的数据库，其数据存储方式有 点类似于Lucene 的 index 文件格式。CouchDB 是一个面向 Web 应用的新一代存储系统。 作为一个分布式的数据库， CouchDB 可以把存储系统分布到n 台物理的节点上面，并且很 好地协调和同步节点之间的数据读写一致性。CouchDB 支持 REST API, 可以让用户使用 JavaScript 来操作 CouchDB 数据库，也可以用 JavaScript 编写查询语句。利用 AJAX 技术结 合 CouchDB 开发出来的 CMS 系统将会非常简单和方便。

CouchDB 是 Couchbase 的子集，不过它提供缓存功能，更好地分片、增量查询、索引 和一些其他的功能。其实 Couchbase 与 CouchDB 也是紧密相关的，Couchbase 产品包含了 CouchDB 的一个副本。

CouchDB 是用 Erlang 开发的面向文档的数据库系统，其数据存储方式类似 Lucene 的 index 文件格式。

**9.4.5** **Neo4j**

大多数的 NoSQL 数据库只是存储键值的一个灵活的捆绑。不过 Neo4j 的存储的是对象 之间的关系，或者说这种结构就是数学中的“图”。Neo4j 是一个面向网络的数据库，也就 是说，它是一个嵌入式的、基于磁盘的、具备完全的事务特性的Java持久化引擎，但是它 将结构化数据存储在网络上而不是表中，当然也可以把 Neo4j 看作是一个高性能的图引擎， 该引擎具有成熟和健壮的数据库的所有特性。该工具包括很多有关搜索和分析的关系的算 法，它能够帮助寻找谁是我的朋友，或者寻找朋友的朋友。这些图的遍历算法，可以节省 很多指针查询的麻烦。

**9.4.6 Oracle 的** **NoSQL**

Oracle 的 NoSQL Database 是 Oracle 开发的产品，将键值对拆分在整个节点集上，这样 的优势在于提供了一个灵活的事务保护措施。从数据在节点上等待存储开始，到通过网络





**大数据基础与应用**

被成功备份结束，都尽在掌握之中。

Oracle 的 NoSQL Database 是 Big Data Appliance 的一个组件，Big Data Appliance 是集 成了Hadoop 、NoSQL Database 、Oracle 数据库 Hadoop 适配器、Oracle 数据库 Hadoop 装载 器及 R 语言的系统。

**9.4.7** **Hadoop的** **HBase**

HBase(Hadoop Database) 是一个高可靠性、高性能、面向列、可伸缩的分布式存储 系统，利用 HBase 可在廉价 PC Server上搭建起大规模结构化存储集群。HBase 是 Google Bigtable的开源应用，类似于Google Bigtable 利用 GFS 作为其文件存储系统，HBase 利用 Hadoop HDFS作为其文件存储系统。Google运行MapReduce 来处理Bigtable中的海量数据， HBase 同样利用 Hadoop MapReduce 来处理 HBase 中的海量数据。

虽然Hadoop 及其所有的工具构成了分布计算开发平台，但 Hadoop 也包括数据库，在 HBase 中也是通过节点来传播数据。Hadoop 的 MapReduce 的架构是非常适合于复杂的计 算任务或查询工作。

**9.4.8** **Bigtable**

Bigtable 是非关系型数据库，是一个稀疏的、分布式的、持久化存储的多维度排序Map。 Bigtable 的设计目的是可靠地处理PB 级别的数据，并且能够部署到上千台机器上。Bigtable 已经实现了下面的几个目标：适用性广泛、可扩展、高性能和高可用性。Bigtable 已经在超 过60个Google 的产品和项目上得到了应用，包括 Google Analytics、Google Finance、Orkut、 Personalized Search 、Writely 和 Google Earth。

**9.4.9** **DynamoDB**

DynamoDB 是亚马逊的键值模式的存储平台，具有良好的可用性和扩展性。读写访问 中99.9%的响应时间都在300ms 内 。DynamoDB 的 NoSQL 解决方案，也是使用键值存储 的模式，并且通过服务器把所有的数据存储在 SSD 上的三个不同的区域。如果有更高的传 输需求， DynamoDB 也可以在后台添加更多的服务器。

**9.4.10** **MongoDB**

MongoDB 是一个介于关系型数据库和非关系型数据库之间的产品，是功能最丰富、最 像关系型数据库的非关系型数据库。它支持的数据结构非常松散，可以存储比较复杂的数 据类型。MongoDB 是一个高性能、开源、模式自由的文档型数据库，使用C++开发，是当 前NoSQL 数据库中最常用的一种数据库。在许多情况下，可以替代传统的关系型数据库， 采用了键值存储方式。

MongoDB 主要解决的是海量数据的访问效率问题。根据资料记载，当数据量达到50GB





第 9 章 NoSQL 数据库

以上的时候，MongoDB 的数据库访问速度是MySQL 的10倍以上。MongoDB 的并发读写 效率不是特别出色，性能测试表明，大约每秒可以处理0.5万～1.5万次读写请求。MongoDB 还自带一个出色的分布式文件系统 GridFS, 可以支持海量的数据存储。MongoDB 也有一 个MongoMapper 项目，是模仿 Merb 的 DataMapper 编写的MongoDB 接口，使用起来非常 简单，几乎和 DataMapper一样，功能非常强大。

**1.** **特点及功能**

MongoDB 最大的特点是支持的查询语言强大，其语法类似于面向对象的查询语言，几 乎可以实现类似关系型数据库单表查询的绝大部分功能，而且还支持对数据建立索引。

(1)面向集合的、模式自由的文档型概念

① 面向集合

文档以分组的方式存储在数据集中， 一组称为一个集合。每个集合在数据库中都有一 个唯一的标识名，并且可以包含无限数目的文档。集合的概念类似于关系型数据库里的表， 不同的是它不需要定义任何模式。

② 模式自由

对于存储在 MongoDB 数据库中的文件，不需要知道它的任何结构定义。对于“无模 式”或“模式自由”概念，从下面例子可以看出，例如下面两条记录可以存在于同一个集 合里面：



{"welcome":"Beijing"}

{"age":25}

③ 文档型

存储的数据是键值对的集合，键是字符串，值可以是数据类型集合里的任意类型，包 括数组和文档。这个数据格式称作 BSON(Binary Serialized document Notation)。

(2)MongoDB 的特点

① 面向集合存储，易于存储对象类型的数据。

② 模式自由。

③ 支持动态查询。

④ 支持完全索引，包含内部对象。

⑤ 文件存储格式为 BSON, 一种 JSON 的扩展。

⑥ 支持复制和故障恢复。

⑦ 使用高效的二进制数据存储，包括大型对象，如视频等。

⑧ 自动处理碎片，以支持云计算层次的扩展性。

⑨ 支持 Python 、PHP 、Ruby 、Java 、C 、C# 、JavaScript 、Perl 及 C++语言的驱动程序。 ⑩ 社区中也提供了对 Erlang 及.NET 等平台的驱动程序。

(3)MongoDB 的功能

① 面向集合的存储：适合存储对象及JSON 形式的数据。

② 动态查询： MongoDB 支持丰富的查询表达式。查询指令使用 JSON 形式的标记，



**大数据基础与应用**

可轻易查询文档中内嵌的对象及数组。

③ 完整的索引支持：包括文档内嵌对象及数组。MongoDB 的查询优化器会分析查询 表达式，并生成一个高效的查询计划。

④ 查询监视： MongoDB 包含一系列监视工具用于分析数据库操作的性能。

⑤ 复制及自动故障转移： MongoDB 数据库支持服务器之间的数据复制，支持主从模 式及服务器之间的相互复制。复制的主要目标是提供冗余及自动故障转移。

⑥ 高效的传统存储方式：支持二进制数据及大型对象(如照片或图片)。

⑦ 自动分片以支持云级别的伸缩性：自动分片功能支持水平的数据库集群，可动态添 加额外的机器。

**2.** **适用场合**

(1)网站数据

MongoDB 非常适合实时插入、更新与查询，并具备网站实时数据存储所需的复制及高 度伸缩性。

(2)缓存数据

由于性能高， MongoDB 也适合作为信息基础设施的缓存层。在系统重启后，由 MongoDB 搭建的持久化缓存层可以避免下层的数据源过载。

(3)大尺寸、低价值的数据

使用传统的关系型数据库存储一些数据时可能会比较昂贵，在此之前，很多时候程序 员往往会选择传统的文件进行存储。

(4)高伸缩性的场景

MongoDB 非常适合由数十或数百台服务器组成的数据库。MongoDB 的路线图中已经 包含对 MapReduce 引擎的内置支持。

(5)用于对象及JSON 数据的存储

MongoDB 的 BSON 数据格式非常适合文档化格式的存储及查询。

**3.** **体系结构**

MongoDB 数据库可以看作一个 MongoDB Server,该 Server 由实例和数据库组成。 一 般情况下， 一个 MongoDB Server 机器上包含一个实例和多个与之对应的数据库，但是在特 殊情况下，如硬件投入成本有限或特殊的应用需求，也允许一个 Server 机器上有多个实例 和多个数据库。

MongoDB 中一系列物理文件(数据文件、日志文件等)的集合或与之对应的逻辑结构 (集合、文档等)称为数据库。简单来说，数据库由一系列与磁盘有关的物理文件组成。

**4.** **数据逻辑结构**

MongoDB 数据逻辑结构是面向用户的，用户使用 MongoDB 开发应用程序使用的就是 逻辑结构。MongoDB 逻辑结构是一种层次结构，如图9-1 所示，主要由文档、集合、数据 库组成。

● MongoDB 的文档相当于关系型数据库中的一条记录；



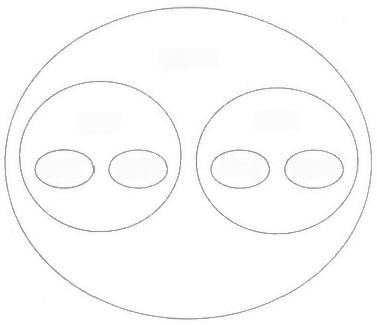


第9章 NoSQL数据库

● 多个文档组成一个集合，集合相当于关系型数据库的表；

● 多个集合逻辑上组织在一起就是数据库；

● 一 个 MongoDB 实例支持多个数据库。

数据库1

集合1 集合2

文档1 文档2 文档1 文档2

图9-1 MongoDB 的数据逻辑结构

MongoDB 与关系型数据库的逻辑结构进行了对比，如表9-6所示。

**表9-6** **MongoDB** **与关系型数据库逻辑结构对比**

|  |  |
| --- | --- |
| MongoDB | 关系型数据库 |
| 文档 | 行 |
| 集合 | 表 |
| 数据库 | 数据库 |

小结

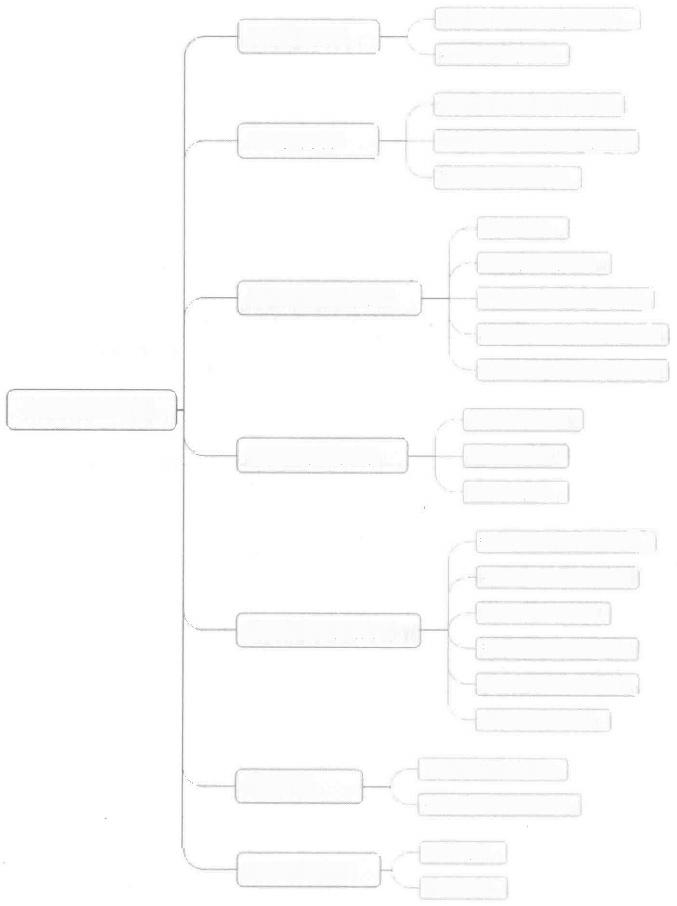
面对越来越多的非结构化数据的存储与管理， NoSQL 数据库应运而生。在大数据中， 85%以上是非结构化数据，因此，对于大数据应用， NoSQL 数据库不可缺少。本章主要介 绍了非结构化问题、NoSQL 的产生背景、特点与优势，在此基础之上，还介绍了 NoSQL 的主要类型，尤其对键值存储方式、文档存储方式、列存储方式等做了较详细的说明。最 后，概括性地介绍了常用的 NoSQL 数据库。



**第10章**

本章主要内容

**大数据预处理技术**



数据抽取的概念与包装器

数据抽取概述

抽取数据的方法

Web数据抽取问题的提出

Web数据抽取的目的与方法

Web数据抽取过程

数据质量

数据清洗的目的

数据清洗算法衡量标准

数据清洗的主要研究成果

数据清洗技术面临的问题

不完整的数据

异常的数据

重复的数据

数据清洗的方法与技巧

数据清洗的实现方式

数据清洗的步骤

数据清洗技术的实现

数据清洗的评价标准

常用的数据清洗算法

大数据清洗工具

数据集成技术概述

数据集成

数据集成系统的构建

数据转换

数据转换与约简

数据约简

数据质量与数据清洗

不符合要求的数据

大数据预处理技术

Web数据抽取

大数据技术主要包括大数据的获取与记录技术，大数据的抽取与清洗技术，大数据集



**第10章** **大数据预处理技术**

成、凝集与表示技术，大数据分析和建模与解释技术等。本章主要介绍大数据抽取、清洗 与集成技术的内容。

**10.1** **数据抽取概述**

大数据来源广泛，数据类型多而繁杂，进而形成了复杂的数据环境，这就给大数据的 获取带来了极大的困难，所以必须对所需要的数据从各种底层数据源进行抽取，从中提取 出关系和实体，经过关联和聚合之后，再采用同一结构来存储这些数据。收集来的数据通 常不能直接用于数据分析，需要将所需要的数据从底层数据源中抽取出来，形成适合于分 析的数据结构。抽取出来的数据不仅包含结构化数据和半结构化数据，也包含图像、视频 等非结构化数据。除此之外，由于监控摄像头、装载有GPS 的智能手机、相机和其他便携 设备无处不在，高保真度的位置和轨迹数据也应在抽取之列。

数据抽取技术的研究异常活跃，主要集中在如下所述几个方面：应用机器学习方法来 增强系统的可移植性、探索更深层次的理解技术、篇幅分析技术、多语言文本处理能力、 Web 信息抽取及对时间信息的处理等。

**10.1.1** **数据抽取的概念与包装器**

**1.** **数据抽取的概念**

数据抽取就是搜索全部数据源，按照某种标准选择合乎要求的数据，并将被选中的数 据传送到目的文件中存储。简单地说，数据抽取过程就是从数据源中抽取数据并传送到目 的文件中的过程。数据源可以是关系型数据库或非关系型数据库，数据源中的数据可以是 结构化数据、非结构化数据和半结构化数据。在数据抽取之前，需要清楚数据源的类型和 数据的类型，以便根据不同的数据源和数据类型采取不同的抽取方法。

数据抽取的定义是：给定数据源 S, 确定一个数据源 S 到数据库的映射 M, 该映射 是从 S 中抽取数据对象，并将这些数据对象按一定格式组装到R 中，这就是数据抽取。 如图10-1所示。



M

S R

图10-1 数据抽取

**2.** **数据抽取包装器**

将完成映射的程序称为包装器，构建包装器需要具备下述知识。

(1)在数据源中抽取数据对象的类型

数据源中的数据对象繁多、千差万别，从简单的字符串到树形结构和有向图结构等， 为了使包装器具有通用性，通常用数据模型来描述数据源中数据对象的结构。



**大数据基础与应用**

(2)在数据源中找到所需的数据对象的方法

可以应用搜索规则驱动一个通用的搜索算法在数据源中搜索与抽取规则相匹配的数 据对象。

(3)为找到的数据选择组装格式

应用符合某个数据库模式的格式来组装已找到的数据对象，对于结构化数据可以利用 关系型数据库格式，对于非结构化数据可以利用文档数据库或键值数据库等格式，对于半 结构化数据可以应用关系型数据库格式和文档数据库或键值数据库结合的格式。

(4)将找到的数据对象组装到数据库中的方法

可以用一组映射规则来描述数据类型与数据库字段之间的关系。当应用抽取算法找到 一个数据对象时，先用映射规则，根据数据对象所属的数据类型找到所对应的数据库字段， 然后将这些数据对象组装到这个字段中， 一般不单独设计组装算法，而是设计完成抽取与 组装功能的算法。

(5)生成和维护数据抽取过程所需的元数据

元数据是指数据抽取模型、抽取规则、数据库模式和映射规则等参数，元数据能够使 抽取和组装算法正常工作。在数据仓库系统中的元数据定义为用于数据仓库管理和有效使 用的任何信息。 一个数据源需要用一套元数据进行描述，由于数据集成系统包含有大量数 据源和元数据，所以维护这些元数据的工作量巨大。

**10.1.2** **抽取数据的方法**

对于不同的数据源，抽取方法不同，常用的几种方法简述如下。

**1.** **同构同质**

如果数据源与组装的目标数据库系统是同构同质，那么 DW 数据库服务器和原业务系 统之间可以在建立直接的链接关系之后，利用结构化查询语言的语句访问。

**2.** **同构异质**

如果数据源与组装的目标数据库系统是同构异质，对于这类数据源可以通过ODBC 的 方式建立数据库链接，例如 Oracle 数据库与 SQL Server 数据库可以建立 ODBC 连接。

**3.** **文件型**

可以利用数据库工具将这些数据导入到指定的数据库，然后从指定的数据库抽取，或 者借助工具完成。

**4.** **增量更新**

对于大数据，需要应用增量抽取。在正常情况下，可以将业务系统记录的业务发生的 时间作为增量标志。在每次抽取之前，先判断 ODL 中记录的最大时间，然后到业务系统中 取出在这个时间内的所有记录。如果业务系统具有时间戳，可以使用时间戳来完成在所设 定时间内所有记录的抽取。





第 1 0 章 大数据预处理技术

**10.2** **Web** **数据抽取**

**10.2.1** **Web 数据抽取问题的提出**

由于Web 数据的复杂性，进而造成 Web 数据抽取出现了一系列困难的问题，主要如下 所述。

**1.Web** **实体的模式构建**

在 Web 数据抽取过程中，需要获取 Web实体的模式信息，以便为进一步识别、抽取来 自不同数据源的 Web 数据对象提供指导。由于 Web 半结构化数据模式具有异构特性和动态 变化的特点，所以有效构建 Web 实体的模式是一个亟待解决的问题。现存的研究方法是利 用搜索引擎搜集尽可能多的网页上的属性标签以及Web 数据对象信息，通过聚类的方法， 一次性生成相关 Web实体的重要属性标签集合。但是，由于 Web 上的半结构化数据模式具 有异构性及动态变化的特点，不同网站上的网页对同类Web 数据对象的描述形式不尽相同， 并且在 Web上将产生新的属性标签，导致该类方法无法满足Web 数据抽取的需要，无法根 据 Web 数据模式的动态变化来实现 Web实体的逐步丰富。

**2.** **包装器的生成**

在 Web 数据抽取过程中，需要准确地从目标网页中抽取目标数据，并对抽取的数据进 行语义理解。通过对目标网站采样页面中的数据元素和属性标签进行准确识别，生成良好 的学习样例，进而生成包装器。现有的大部分方法主要利用采样页面自身包含的特征，例 如视觉特征、结构特征等，识别出页面中的数据元素和属性标签，生成学习样例。但是， 对于结构复杂的 Web 页面，利用上述方法产生的学习样例质量不高，严重地影响了生成的 包装器对目标数据抽取的准确度。

**3.** **丰富** **Web** **实体模型**

在 Web数据抽取过程中，需要建立新发现的Web 实体与已有的 Web 实体之间的联系， 进而丰富模型，为进一步整合 Web 数据奠定基础。

**10.2.2** **Web 数据抽取的目的与方法**

**1.Web 数据抽取的目的**

在 Internet 中的数据多以HTML 页面形式存于各个网站。但是 HTML 描述语言主要是 描述数据的显示，而不是用于描述数据本身，所以 HTML 没有结构限制和数据类型定义， 而且HTML 页面中的半结构化数据和无结构化数据，无法直接应用。Web 数据抽取的目的 是从半结构化数据和无结构化数据中抽取特定的事实数据，例如从产品销售网站上抽取产





**大数据基础与应用**

品的名称、规格、价格和数量等数据，从股票行情网站上抽取股票代码、名称、价格、交 易量等数据，从气象网站上抽取城市代码、地名、温度、湿度、风力、风向以及未来几天 的天气预报等数据。将这些抽取出来的数据以结构化方式表示之后，就可以很容易存于数 据库、XML 文件中。

**2.Web 数据抽取的方法**

(1)Web 数据抽取技术的基本内容

Web 数据抽取技术主要包括网页获取、抽取方法、抽取规则、数据校验与数据集成等 内容。技术的核心是从半结构化或非结构化数据中，识别出目标系统需要的数据，并将其 转为结构化的、语义清晰的格式。

(2)Web 数据抽取模型

为了抽取特定网站上的 Web数据，需要使用某种方法构建一个基于该数据源的适用模 型，针对一组页面进行处理，通常将这种模型称为包装器。包装器由抽取规则和抽取器两 部分组成，抽取规则用于描述抽取规则，包括网页结构、数据项定位、抽取步骤、转换、 校验规则、输出格式等。抽取器用于执行基于上述数据抽取规则建立的执行程序或其他应 用程序，显然，抽取规则是 Web 数据抽取技术的核心内容。

一个完整的 Web 数据抽取系统主要由包装器、网页下载、超链接分析、预处理、数据 完整性校验、数据映射、数据组装等组成。

**10.2.3** **Web 数据抽取过程**

Web 数据抽取过程一般包括五个步骤，即网页获取、数据抽取、数据校验与转换、数 据存储与数据组装。

**1.** **网页获取**

网页分为两类， 一类是仅包含需要抽取数据的数据网页，另一类是导航网页。在导航 网页中，包含指向数据网页的若干链接。在很多情况下，数据网页也包括导航信息，将两 类网页在物理上组合在一起。

网页获取是指通过指定一组 Web 页面的网址，或者根据指定的超链接导航规则，由抽 取程序自动爬行获取网页。当被访问的页面有权限保护或限制时，需要执行脚本程序登录 或通过认证之后访问页面。当网站不仅使用静态链接，而且还使用了Web 表单或 JavaScript 代码的形式实现链接导航，对于这种导航网页，需要手工分析页面代码并设计相应的处理 程序。

**2.** **数据抽取**

当获取网页之后，进行数据抽取。 一般使用包装器完成数据抽取。主要的任务是根据 抽取规则对 Web 页面进行解析处理，抽取需要的数据项。在抽取之前，需要对HTML 文件 预处理，例如对HTML 进行规范化和纠错等。





**第10章** **大数据预处理技术**

**3.** **数据校验与转换**

对上一步抽取的数据，需要进行数据校验与转换。数据校验能够保证抽取数据的正确 性和完整性。对于明显的错误可以使用正则表达式、校验规则等手段加以限制，并应用特 定的领域知识和机器学习等技术纠错或将丢失的数据补充完整。数据转换能够保证数据的 一致性。当从多个网站的数据源抽取并集成数据时，不同的网站可能使用不同的命名规范、 不同的表示形式和不同的计量单位，即使同一个网站也可能出现这种不一致现象。此时可 以根据数据字典，将数据映射到统一的标准格式来保证数据质量，具体的实现方法是应用 领域知识、使用条件声明、规则表达式等。

**4.** **数据存储**

抽取到的数据可以根据应用系统的需求，采用不同的方式进行存储。关系形式的数据 可以存于关系型数据库中。如果需要跨平台共享，可以使用 XML 存储。如果需要以面向 对象的形式存储数据，可以用面向对象数据库的形式存储。

**5.** **数据组装**

抽取的数据来自多个网站，由于页面格式不一致，数据组成也不同，即使是同一个站 点，也存在一个逻辑数据被分割到多个页面的情况，此时应该进行数据组装。此外，页面 中的数据一般是后台结构化数据的一个视图。

**10.3** **数据质量与数据清洗**

**10.3.1** **数据质量**

数据是进行数据分析的最基本资源，高质量的数据是保证完成数据分析的基础。尤其 是大数据具有数据量巨大、数据类型繁多和非结构化等特征，为了快速而准确地获得分析 结果，提供高质量的大数据尤其重要。数据质量与绩效之间存在着直接关联，高质量的数 据可以满足需求，有益于获得更大价值。

通常从准确性、完整性、 一致性和及时性四个方面来描述数据质量，这四个方面又称 为数据质量的四要素。

**1.** **数据的准确性**

数据的准确性是对数据真实性的描述，是对所存储数据的准确程度的描述。数据不准 确的表现形式是多样的，例如字符型数据的乱码现象、异常大或者异常小的数值、不符合 有效性要求的数值等。由于很难发现没有明显异常错误的数据，所以对数据准确性的监测 也是困难的问题。

**2.** **数据的完整性**

数据完整性是对数据记录的完整性、记录是否存在缺失的描述。例如数据的缺失可以





**大数据基础与应用**

是记录的缺失或记录中某个字段的缺失，两者都可能造成统计结果的不准确。所以完整性 是数据质量最基础的保障，而对完整性的检测也比较容易。

**3.** **数据的一致性**

数据的一致性主要包括数据记录规范的一致性和数据逻辑的一致性。 (1)数据记录规范的一致性

数据记录规范的一致性主要是指数据编码和格式的问题，例如网站的用户ID 是 1 5 位 的数字，商品 ID 是10位数字，商品包括20个类目， IP 地址一定是用“.”分隔的由4个 0～255的数字组成，遵循确定的规范，所定义的数据也遵循确定的约束，如完整性的非空 约束、唯一值约束等。这些规范与约束使得对于数据记录有统一的规范，进而保证了数据 记录的一致性。

(2)数据逻辑的一致性

数据逻辑的一致性主要是指标统计和计算的一致性，比如 PV>=UV, 新用户比例在0~ 1 之间等。在数据质量中，数据的一致性是比较重要也是比较复杂的内容。

**4.** **数据的及时性**

数据从产生到可以查看的时间间隔称为数据的延时时间。虽然分析数据的实时性要求 并不是太高，但并不表明就没有要求。如果数据要延时两三天才能出来，或者每周的数据 分析结果需要两周后才能出来，那么分析的结论可能已经失去时效性。如果某些实时分析 和决策需要用到小时或者分钟级的数据，这时的需求对数据的时效性要求就更高。所以， 及时性也是数据质量的重要因素之一。

**10.3.2** **数据清洗的目的**

数据清洗的目的是为了获得高质量的数据，进而提高数据的可利用性，主要包括消除 异常数据、清除重复数据、保证数据的完整性等。数据清洗就是指通过分析脏数据产生的 原因和存在形式，构建数据清洗的模型和算法完成脏数据的清除，进而实现将不符合要求 的数据转化成满足数据质量或应用要求的数据，提高数据集的数据质量。

**10.3.3** **数据清洗算法衡量标准**

数据清洗是一项与领域密切相关的工作，由于各领域的数据质量不一致、充满复杂性， 所以还没有形成通用的国际标准，只能根据不同的领域制定不同的清洗算法。从目前来看， 数据清洗算法的衡量标准主要有下述几方面。

**1.** **返回率**

返回率是指重复数据被正确识别的百分率。

**2.** **错误返回率**

错误返回率是指错误数据占总数据记录的百分比。



第10章 大数据预处理技术

**3.** **精确度**

精确度是指算法识别出的重复记录中的正确的重复记录所占的百分比，计算方法如下： 精确度=100%-错误返回率

**10.3.4** **数据清洗的主要研究成果**

目前，数据清洗的主要研究成果如下。

● 提出高效的数据异常检测算法，避免扫描整个大型系统；

● 在自动检测数据异常和进行清洗处理的过程中，增加交互处理，可以防止对正确 数据的错误处理；

● 对数据集文件的处理；

● 消除合并后数据集中的重复数据；

● 建立一个通用的、与领域无关的数据清洗框架；

● 模式集成问题。

目前已经提出的数据清洗的模型主要有：基于聚类模式的数据清洗模型、基于粗糙集 理论的数据清洗模型、基于模糊匹配的数据清洗模型、基于遗传神经网络的数据清洗模型 和基于专家系统的数据清洗模型等。

虽然利用这些模型可以完成不同程度的数据清洗，但是都存在一些不足和弊端。例如， 聚类模式的数据清洗模型直接检测异常数据作用不大，主要缺陷是耗时，特别是在记录条 数多时不适合于检测异常数据。

在运用聚类算法的基础之上，使用给予模式的方法，即每个字段使用欧式距离，类别 K-mean 算法，仅检测到较少数的记录(约30%)满足超过90%字段的模式。

经典的关联规则难以发现异常，但数量型关联规则、序数规则能够较好地检测异常与错误。

数据清洗技术的研究还处于初级阶段，虽然数据清洗异常重要，但并没有引起足够重 视。数据清洗的通用框架的研究、准确地识别重复记录的研究等大量工作需要去做。因为 互联网已经成为大数据的主要来源，所以对 Web 搜索引擎返回的结果进行清洗是一个必须 考虑的问题。随着 XML 数据处理标准的日益成熟， XML 文档的定义和对 XML 文档的数 据清洗的研究已呈现在我们面前。

**10.3.5** **数据清洗技术面临的问题**

数据清洗技术的研究主要面临下述问题。

● 目前数据清洗主要应用在对客户数据要求较高的特定领域，例如银行、保险和证 券领域，针对本行业的具体应用开发软件，缺少通用的数据清洗框架研究。

● 清洗工具较少，尤其是对于不完整数据和异常数据的清洗研究较少。

● 对于重复数据的检测需要投入更大的研究，这是由于面对大数据时，需要提供耗 时小、检测精度高、效率高的算法。

● 目前的ETL 工具的研究集中在数据仓库中的数据清洗，但是对于Web 数据清洗工 具研究较少。



**大数据基础与应用**

由于中文与英文的语言特点不同，所以清洗的方法也不同，但对中文的数据清洗 应引起重视并加强研究。

**10.4** **不符合要求的数据**

归纳起来，数据清洗是指清洗掉不符合要求的数据，主要包括三大类：不完整的数据、 异常的数据、重复的数据。在大数据清洗中，应用了MapReduce 分布编程模型，由计算机 程序完成不符合要求的数据的清洗。

**10.4.1** **不完整的数据**

不完整的数据是指一些应该有的信息缺失，例如供应商的名称、分公司的名称、客户的 区域信息缺失，业务系统中主表与明细表不能匹配等。将这一类数据过滤出来，按缺失的内 容分别写入不同数据文件向客户提交，要求在规定的时间内补全。补全之后写入数据仓库。

由于编码和录入误差，数据中可能存在一些无效值和缺失值，需要给予适当的处理。 常用的处理方法有：估算、整列删除、变量删除和成对删除。

最简单的办法就是用某个变量的样本均值、中位数或众数代替无效值和缺失值。这种 办法简单，但没有充分考虑数据中已有的信息，误差可能较大。另一种办法就是根据调查 对象对其他问题的答案，通过变量之间的相关分析或逻辑推论进行估计。例如，某一产品的 拥有情况可能与家庭收入有关，可以根据调查对象的家庭收入推算拥有这一产品的可能性。

整列删除是剔除含有缺失值的样本。这种做法的结果可能导致有效样本量大大减少， 无法充分利用已经采集到的数据。因此，只适合关键变量缺失，或者含有无效值、缺失值 的样本比重很小的情况。

如果某一变量的无效值和缺失值很多，而且该变量对于所研究的问题不是特别重要， 则可以考虑将该变量删除。这种做法减少了分析所用的变量数目，但没有改变样本量。

成对删除是用一个特殊码(通常是9、99、999等)代表无效值和缺失值，同时保留数 据集中的全部变量和样本。但是，在具体计算时只采用有完整答案的样本，因而不同的分 析因涉及的变量不同，其有效样本量也会有所不同。这是一种保守的处理方法，最大限度 地保留了数据集中的可用信息。

采用不同的处理方法可能对分析结果产生影响，尤其是当缺失值的出现并非随机，且变 量之间明显相关时。因此，在调查中应当尽量避免出现无效值和缺失值，保证数据的完整性。

**10.4.2** **异常的数据**

**1.** **一致性检查**

一致性检查是根据每个变量的合理取值范围和相互关系，检查数据是否合乎要求，发 现超出正常范围、逻辑上不合理或者相互矛盾的数据。例如，用1～7级量表测量的变量出



**第10章** **大数据预处理技术**

现了0值，体重出现了负数，都应视为超出正常值域范围。SPSS 、SAS 和 Excel 等计算机 软件都能够根据定义的取值范围，自动识别每个超出范围的变量值。具有逻辑上不一致性 的答案可能以多种形式出现。例如，许多调查对象说自己开车上班，又报告没有汽车；或 者调查对象报告自己是某品牌的重复购买者和使用者，但同时又在熟悉程度量表上给了很 低的分值。发现不一致时，要列出问卷序号、记录序号、变量名称、错误类别等，便于进 一步核对和纠正。

**2.** **错误的数据**

错误的数据产生的原因是系统不够健全，在接收输入后没有进行判断就直接写入后 台数据库，如数值数据输入为全角数字字符、字符串数据后面有一个回车、日期格式不 正确、日期越界等。错误的数据也要进行分类，对于类似于全角字符、数据前后有不可 见字符的问题，只能通过写查询的方式找出来，然后要求客户在业务系统修正之后抽取。 日期格式不正确或者日期越界这类错误会导致系统运行失败，对于这类错误，需要去业 务系统数据库用查询的方式挑出来，交给业务主管部门，要求限期修正，修正之后再进 行抽取。

**10.4.3** **重复的数据**

重复的数据是指冗余数据，对于这类数据应删除，具体的工作如下。

● 空值填充；

● 统一单位；

● 标准化处理；

● 删除无必要的变量；

● 逻辑值是否有错误检查；

● 需要引入新的计算变量；

● 需要排序；

● 进行主成分或者因子分析。

**10.5** **数据清洗技术的实现**

数据清洗遵循下述4个原则：格式化、主键约束、数据不缺、数据检测。

数据清洗是一个反复的过程，不可能在几天内完成，只有不断地发现问题、解决问 题才能够完成。对于是否过滤，是否修正， 一般要求客户确认。对于过滤掉的数据，写 入数据文件或者将过滤数据写入数据表，在 ETL 开发的初期可以每天向业务单位发送过 滤数据的邮件，促使他们尽快地修正错误，同时也可以作为将来验证数据的依据。数据 清洗需要注意的是不要将有用的数据过滤掉，依据每个过滤规则认真进行验证，并由用 户确认。





**大数据基础与应用**

**10.5.1** **数据清洗的方法与技巧**

**1.** **使用断言**

断言是指轻易地在短时间内对某一件事情或者事物下的一种主观性非常强的言论。在 数据清洗过程中，可以使用断言找出程序代码中的 bug。用断言的形式描述代码格式的假 设，如果一旦发现有数据与断言相悖，就修改这些断言。例如，如果记录是有序的，则断 言每一条记录都是有7个字段吗?如果是，则断言每一个字段都是0～26之间的奇数吗? 如果是，则断言……总之，能断言的都断言。

在理想世界中，所有记录都应该是规整的格式，并且遵循某种简洁的内在结构。但是 实际中并不是这样，所以书写断言并非容易之事。清洗数据的程序会经常崩溃，因为每一 次崩溃都意味着这些数据又与最初的假设相悖了。反复地改进断言直到能成功地通过。但 一定要尽可能让它们保持严格，否则，可能达不到所需的效果。最坏的情况不是程序运行 不通，而是运行出来不是所需的结果。

**2.** **谨慎处理跳过的记录**

原始数据中有些记录是不完整或者是损坏的，所以清洗数据的程序只能跳过。简单地 跳过这些记录不是最好的办法，因为不知道什么数据被遗漏了。因此，建议如下：

① 打印出 warning 提示信息，这样就能够在记住之后，再去寻找什么地方出错。

② 总计跳过了多少记录，成功清洗了多少记录，这样做能够对原始数据的质量有个大 致的了解。例如，如果只跳过了0.5%,这还说得过去，但是如果跳过了35%,那就需要检 查这些数据或者代码存在什么问题。

**3.** **存储变量的类别以及类别出现的频次**

将变量的类别以及类别出现的频次存储起来可以提高清洗效率。数据中经常有些字段 是枚举类型的。例如，血型只能是A 、B 、AB 或者 O。用断言来限定血型只能是这4种之 一，但是如果某个类别包含多种可能的值，就不能用断言了。这时可以采用counter 这种数 据结构来存储，做如下处理：

① 对于某个类别，如果碰到了始料未及的新取值时，就能够打印一条提醒消息。

② 清洗完数据之后，可进行检查。例如，假如有人把血型误填成 C,通过检查就可以 轻松发现。

**4.** **断点处继续清洗**

如果有大量的原始数据需要清洗，要一次清洗完可能需要很长时间，多达几天。实际 上，经常在洗到一半的时候突然崩溃了。假设有100万条记录，清洗程序在第325392条因 为某些异常崩溃了，修改了这个 bug 之后，然后重新清洗，这样的话，程序就得重新从1 清洗到325391,这是在做无用功。为了克服上述问题，可以做如下处理：

① 应用清洗程序打印出当前正在清洗第几条，如果崩溃了，就能知道处理到哪条时崩

**第10章** **大数据预处理技术**

溃了。

② 使用程序支持在断点处开始清洗，这样当重新清洗时，就能从第325392条直接开始。 重洗的代码有可能会再次崩溃，只要再次修正bug, 然后从再次崩溃的记录开始就行了。

当所有记录都清洗结束之后，再重新清洗一遍，因为后来修改 bug 后的代码可能会对 之前的记录的清洗带来一些变化，两次清洗保证万无一失。但总的来说，设置断点能够节 省很多时间，尤其是在调试的时候，效果更为明显。

**5.** **在部分数据上进行测试**

不要希望一次性清洗所有数据。当刚开始写清洗代码和调试的时候，在一个规模较小 的子集上进行测试，然后扩大测试的这个子集再测试。这样做的目的是能够使清洗程序很 快地完成测试集上的清洗，少则几秒，将节省反复测试的时间。但是这样做时要注意，用 于测试的子集往往不能涵盖到一些比较少见的奇异记录。

**6.** **把清洗日志打印到文件中**

当运行清洗程序时，把清洗日志和错误提示都打印到文件中，这样就能轻松地使用文 本编辑器来查看。

**7.** **把原始数据一并存储下来**

当不考虑存储空间的时候，这一条经验还是很有用的。这样做能够让原始数据作为一 个字段保存在清洗后的数据中，在清洗完之后，如果发现哪条记录存在问题了，就能够直 接看到原始数据是什么样子，方便调试。不过，这样做的坏处就是需要消耗双倍的存储空 间，并且让某些清洗操作变得更慢，所以这一条只适用于效率允许的情况下。

**8.** **验证清洗后的数据**

需要编写一个验证程序来验证清洗后得到的数据是否跟预期的格式一致。不能控制原 始数据的格式，但是能够控制清洗后数据的格式。所以， 一定要确保干净数据的格式是符 合预期的格式的。这一点其实是非常重要的，因为完成了数据清洗之后，接下来就会直接 在这些干净数据上进行下一步工作了。除非万不得已，甚至再也不会碰那些原始数据了。 因此，在开始数据分析之前要确保数据是足够干净的。否则，可能会得到错误的分析结果， 到那时候，就很难再发现很久之前的数据清洗过程中所犯的错误了。

**9.** **解决不完整数据的方法**

大多数情况下，缺失的值必须重新填入。某些缺失值可以从本数据源或其他数据源推 导出来，这就可以用平均值、最大值、最小值或更为复杂的概率估计代替缺失的值，从而 达到清理的目的。

**10.** **错误值的检测及解决方法**

用统计分析的方法识别可能的错误值或异常值，如偏差分析、识别不遵守分布或回归 方程的值，也可以用简单规则库(常识性规则、业务特定规则等)检查数据值，或使用不





**大数据基础与应用**

同属性间的约束、外部的数据来检测和清理数据。

**11.** **不一致性(数据源内部及数据源之间)的检测及解决方法**

不同的企业有不同的业务规则，所不同的技术指标应转化为同一的粒度。

从多数据源集成的数据可能有语义冲突，可定义完整性约束用于检测不一致性，也可 通过分析数据发现联系，从而使得数据保持一致。数据不一致包括同一数据源的不一致和 多个数据源之间的数据不一致等类别，例如在一个应用系统中，BJ 表示北京，SH 表示上 海， GZ 表示广州，而在另一个应用系统中，对应的代码为1、2、3。

**12.** **重复记录的检测及消除方法**

数据库中属性值相同的记录被认为是重复记录，通过判断记录间的属性值是否相等来 检测记录是否相等，相等的记录合并为一条记录。合并、清除是消重的基本方法。

大数据时代，数据规模庞大、增长迅速、类型繁多、非结构化已成为无法回避的现实 问题。如何把繁杂的大数据变成我们能处理的、有效的小数据，即针对特定问题而构建一 个干净、完备的数据集，这一过程变得尤为重要。在大数据时代，如果不加强数据清洗，

则垃圾进、垃圾出的现象会更加严重。

**10.5.2** **数据清洗的实现方式**

通过编写专门应用程序的方法能解决某个特定的问题，但不够灵活，特别是清理过程 需要反复进行。 一般来说，很少有数据清洗一遍就达到要求，将导致程序复杂，清理过程 变化，工作量大。这种方法没有充分利用数据库提供的强大数据处理能力。

能够解决某类特定应用域的问题，如根据概率统计学原理查找数值异常的记录，对姓 名、地址、邮政编码等进行清理，这是目前研究得较多的领域，也是应用最成功的一类。

例如商用系统： Trillinm Software 、System Match Maker 等。

与特定应用领域无关的数据清理，这一部分的研究主要集中在清理重复的记录上，如 Data Cleanser 、Data Blade Module 、Integrity 系统等。

绝大部分数据清理方案提供接口用于编制清理程序， 一般来说包括很多耗时的排序、 比较、匹配过程，且这些过程多次重复，用户必须等待较长时间。在一个交互式的数据清 理方案中，系统将错误检测与清理紧密结合起来，用户能通过直观的图形化界面一步步地 指定清理操作，且能立即看到此时的清理结果(仅仅在所见的数据上进行清理，所以速度 很快),不满意清理效果时还能撤销上一步的操作，最后将所有清理操作编译执行。这种方 案对清理循环错误非常有效。

许多数据清理工具提供了描述性语言解决用户友好性，降低用户编程复杂度。如 ARKTOS 方案提供了XADL 语言(一种基于预定义的 DTD 的 XML 语言)、SADL 语言， 在 ATDX 提供了一套宏操作(来自于 SQL 语句及外部函数),一种 SQL2Like 命令语言， 这些描述性语言都在一定程度上减轻了用户的编程难度，但各系统一般不具有互操作性， 不能通用。





**第10章** **大数据预处理技术**

数据清理属于一个较新的研究领域，直接针对这方面的研究并不多，中文数据清理更 少。现在的研究主要为解决两个问题：发现异常、清理重复记录。

**10.5.3** **数据清洗的步骤**

从通用性和实用性角度出发，提出了多种数据清洗方法。但是不管哪种方法，基本都 由三个阶段组成：数据分析，定义错误类型；搜索、识别错误记录；修正错误。

**1.** **定义和确定错误的类型**

(1)分析数据

分析数据是数据清洗的前提与基础，通过详尽的分析数据来检测数据中的错误或不一 致情况，可以使用分析程序来获得关于数据属性的元数据，从而发现数据集中存在的质量 问题。

(2)定义数据清洗转换规则

根据上一步进行数据分析得到的结果来定义清洗转换规则与工作流程。根据数据源的 个数，数据源中不一致数据和脏数据多少的程度，需要执行大量的数据转换和清洗操作。 运用 MapReduce 分布编程模型，使转换代码的自动生成变成可能。

**2.** **搜寻并识别错误的实例**

(1)自动检测属性错误

检测数据集中的属性错误，需要花费大量的人力、物力和时间，而且这个过程本身很 容易出错，所以需要利用高级的方法自动检测数据集中的属性错误，方法主要有：基于统 计的方法、聚类方法、关联规则的方法。

(2)检测重复记录的算法

消除重复记录可以针对两个数据集或者一个合并后的数据集，首先需要检测出标识同 一个现实实体的重复记录，即匹配过程。检测重复记录的算法主要有：基本的字段匹配算 法、递归的字段匹配算法、Smith-Waterman 算法和 Cosine 相似度函数等。

**3.** **纠正所发现的错误**

在数据源上执行预先定义好的并且已经得到验证的清洗转换规则和工作流。当直接在 数据源上进行清洗时，需要备份数据源，以防需要撤销上一次或上几次的清洗操作。清洗 时根据脏数据存在形式的不同，执行一系列的转换步骤来解决模式层和实例层的数据质量 问题。为处理单数据源问题并且为其与其他数据源的合并做好准备， 一般在各个数据源上 应该分别进行几种类型的转换，主要包括：

(1)从自由格式的属性字段中抽取值(属性分离)

自由格式的属性一般包含很多信息，而这些信息有时需要细化成多个属性，从而进一 步支持后面重复记录的清洗。

(2)确认和改正

自动处理输入和拼写错误，基于字典查询的拼写检查对于发现拼写错误是很有用的。



**大数据基础与应用**

(3)标准化

为了使记录实例匹配和合并变得更方便，应该把属性值转换成一致和统一的格式。 (4)干净数据回流

当数据被清洗后，干净的数据应该替换数据源中原来的脏数据。这样可以提高原系统 的数据质量，还可避免将来再次抽取数据后进行重复的清洗工作。

**10.5.4** **数据清洗的评价标准**

**1.** **数据的可信性**

可信性包括精确性、完整性、 一致性、有效性、唯一性等指标。

① 精确性：描述数据是否与其对应的客观实体的特征相一致。

② 完整性：描述数据是否存在缺失记录或缺失字段。

③ 一致性：描述同一实体的同一属性的值在不同的系统是否一致。

④ 有效性：描述数据是否满足用户定义的条件或在一定的域值范围内。 ⑤ 唯一性：描述数据是否存在重复记录。

**2.** **数据的可用性**

数据的可用性考察指标主要包括时间性和稳定性。

① 时间性：描述数据是当前数据还是历史数据。

② 稳定性：描述数据是否稳定，是否在其有效期内。

**3.** **数据清洗的代价**

数据清洗的代价即成本效益，在进行数据清洗之前考虑成本效益这个因素是很必要的。 因为数据清洗是一项十分繁重的工作，需要投入大量的时间、人力和物力。在进行数据清 洗之前要考虑其物质和时间的开销，是否会超过组织的承受能力。通常情况下大数据集的 数据清洗是一个系统性的工作，需要多方配合以及大量人员的参与，需要多种资源的支持。

企业所做出的每项决定目标都是为了获得更大的经济效益，如果花费大量资金、时间、 人力和物力进行大规模的数据清洗之后，所能带来的效益远远低于投入，那么这样的数据 清洗被认定为一次失败的数据清洗，故在进行数据清洗之前进行成本效益的估算是非常重 要的。

**10.5.5** **常用的数据清洗算法**

**1.** **空缺值的清洗**

对于空缺值的清洗可以应用软件自动填写空缺值，可以使用一个全局变量填充空缺值， 使用属性的平均值、中间值、最大值、最小值或更为复杂的概率统计函数值来填充空缺值。



**第10章** **大数据预处理技术**

**2.** **噪声数据的清洗**

分箱 (Binning), 通过考察属性值的周围值来平滑属性的值：属性值被分布到一些等 深或等宽的“箱”中，用箱中属性值的平均值或中值来替换“箱”中的属性值；计算机检 测可疑数据；使用简单规则库检测和修正错误；使用不同属性间的约束检测和修正错误；

使用外部数据源检测和修正错误。

**3.** **不一致数据的清洗**

对于有些事务，所记录的数据可能存在不一致。有些数据不一致可以加以更正。例如， 知识工程工具可以用来检测违反限制的数据。例如，知道属性间的函数依赖，可以查找违 反函数依赖的值。此外，数据集成也可能产生数据不一致。

**4.** **重复数据的清洗**

目前消除重复记录的基本思想是“排序和合并”,先将数据库中的记录排序，然后通过 比较邻近记录是否相似来检测记录是否重复。

消除重复记录的算法主要有：优先队列算法、近邻排序算法、多趟近邻排序算法等。

**10.5.6** **大数据清洗工具**

目前开发的数据清理工具大致可分为三类：数据迁移工具、数据清洗工具、数据审计 工具。

数据迁移工具允许指定简单的转换规则，如将字符串 gender 替换成 sex 。sex 公司的 PrismWarehouse 是一个流行的工具，就属于这类。

数据清洗工具使用领域特有的知识对数据做清洗，它们通常采用语法分析和模糊匹配 技术完成对多数据源数据的清理。某些工具可以指明源的相对清洁程度。工具 Integrity 和 Trillum 属于这一类。

数据审计工具可以通过扫描数据发现规律和联系。因此，这类工具可以看作是数据挖 掘工具的变形。

**10.6** **数** **据** **集** **成**

数据集成是指把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中，从 而提供全面的数据共享，建立数据仓库的过程就是数据集成。主要考虑实体识别、数据冗 余和数据值冲突检测与处理。在数据集成领域，已经有了很多成熟的框架可以利用。目前 通常采用联邦式、基于中间件模型和数据仓库等方法来构造集成的系统，这些技术着重解 决数据共享和提供决策支持的问题。





**大数据基础与应用**

**10.6.1** **数据集成技术概述**

由于开发信息系统的时间或开发部门的不同，往往有多个异构的、运行在不同的软硬 件平台上的信息系统同时运行，这些系统的数据源彼此独立、相互封闭，使得数据难以在 系统之间交流、共享和融合，从而形成了信息孤岛。随着信息化应用的不断深入，企业内 部、企业与外部信息交互的需求日益强烈，急切需要对已有的信息进行整合，连通信息孤 岛，共享信息。正是在这一背景下，数据集成技术出现，并得以迅速发展。

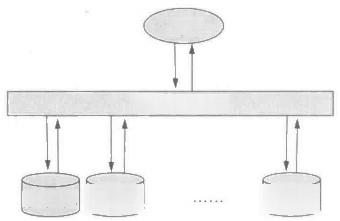
数据集成时，主要考虑如下三个问题：①来自多个信息源的现实实体的匹配问题，即 来自不同的数据源的实体是否是同一实体，完成实体识别需要元数据帮助；②利用相关分 析来检测冗余与重复；③同一实体在不同数据源的属性值不同，需要检测与处理。

**1.** **数据集成技术的任务**

数据集成技术的任务是将相互关联的分布式异构数据源集成到一起，使用户能够以透 明的方式访问这些数据源。在这里，集成是指维护数据源整体上的数据一致性，提高信息 共享利用的效率，透明方式是指用户不必关心如何对异构数据源进行访问，只关心用何种 方式访问何种数据。

**2.** **数据集成系统**

实现数据集成的系统称为数据集成系统，它能够为用户提供统一的数据源访问接口， 执行用户对数据源访问的请求，如图10-2所示。



用户

数据集成系统

数据源

数据源

数据源1

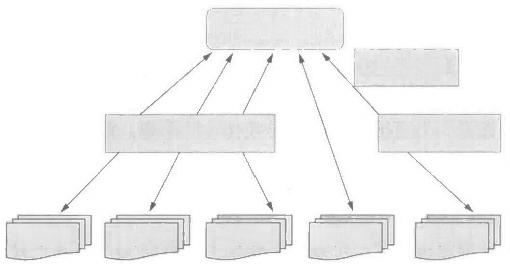
图10-2 数据集成系统

通过应用间的数据交换从而达到数据集成，数据集成主要解决数据的分布性和异构性 的问题。数据集成能够为各种异构数据提供统一的标识、存储和管理，屏蔽了各种异构数 据间的差异，通过异构数据集成系统完成统一操作，因此，只有异构数据对用户来说是无 区别的。数据集成将存于自治和异构数据源中的数据进行组合，可向用户提供一个统一的 模式，用于用户提交查询。其中典型的是全局模式，又称中介模式。数据集成系统必须预 先建立全局模式与数据源之间的语义映射。利用语义映射，用户提交的基于全局模式的查 询转换成对于各数据源的可执行的一系列查询，其架构如图10-3所示。





第10章 大数据预处理技术

全局模式

查询重写

语义映射

优化与执行

包装器 包装器 包装器 包装器 包装器

图10-3 数据集成系统的架构

(1)全局模式

全局模式通过提供一个统一的数据逻辑视图来隐藏底部的数据细节，进而可以使用户 将集成的数据源看成一个统一的整体。数据集成系统通过全局模式将各个数据源的数据集 成，但被集成的数据仍存储在各局部数据源中，通过各数据源的包装器对数据进行转换， 将数据转换成全局模式。由于用户的查询是基于全局的查询，不用知道每个数据源的模式， 即每个数据源的模式对用户透明。中介器将基于全局模式的一个查询转换为基于各局部数 据源模式的一系列查询，交给查询引擎优化与执行。对每个数据源的查询都返回结果数据， 中介器对这些数据连接与集成，最后将符合用户查询要求的数据返回给用户。

全局模式还解决了各数据源中的数据更新问题。当底层数据源发生变化时，只需修改 全局模式的虚拟逻辑图，较显著地减少了系统的维护开销。与数据仓库相比较，优势更为 明显。

在数据更新时，数据仓库的处理方法更为复杂。其过程是必须将各数据源的所有数据 都预先取到一个中心仓库中，当数据发生变化时，还需要到底层数据源中再取一次，还要 更新与这些变化了的数据相关的那些数据，维护开销显然增大。

(2)语义映射

这里所介绍的映射是全局模式与数据源模式之间的映射，它将多个数据源模式映射到 全局模式上。常用的数据集成技术中的映射关系主要有下述两种。

① 全局视图映射法。全局视图映射法是将本地数据源的局部视图映射到全局视图，也 就是说，全局视图可以被描述成源模式上的一组视图。用户可以直接查询数据源模式上的 全局视图。本方法的优点是查询效率高，缺点是构造出的映射关系可扩展性差，不适合于 数据源变化的情况。这是因为当局部数据源发生变化时，全局视图必须进行修改，维护困 难，开销大。

② 局部视图映射法。局部视图映射法是将全局视图映射到各数据源上的本地局部视 图，即将各数据源模式描述为全局模式上的视图。当用户提交某个查询时，中介系统通过 整合不同的数据源视图决定如何应答查询。本方法的优点是映射关系的可扩展性好，适合 于信息源变化较大的情况，缺点是查询效率低，容易丢失信息。



**大数据基础与应用**

**3.** **查询重写**

数据集成系统为多数据源提供了统一的接口，利用视图描述了一个自治的、异构的数 据源的集合。用户基于全局模式提交一个查询，数据集成系统通过源模式与全局模式之间 的映射关系将该查询重写为可接受的语法形式传给数据源，此后，基于数据源的查询被优 化并执行。

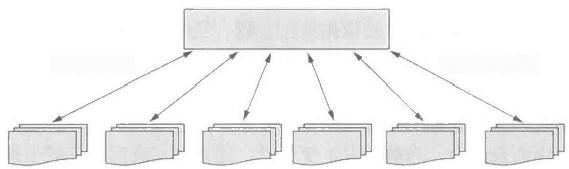
**4.** **利用视图应答查询**

利用视图应答查询是指给定一个数据库模式上的查询 Q, 与同一数据库模式上的视图 定义集 V={V₁,V₂,…Vn}, 能够使用视图V₁ 、V₂……Vm获得对查询Q 的应答。

**10.6.2** **数据集成系统的构建**

**1.** **模式之间映射关系的生成**

模式之间的语义映射关系是构建数据集成系统的基础，更准确地说，建立源模式和全 局模式之间的语义关系。其方法是：使用机器学习的方法，首先建立一个初步的模式映射 关系，并将这些关系作为学习数据，然后对这些映射归纳，再预测出其他未知的模式之间 的映射关系，基于机器学习方法的全局模式与源模式映射如图10-4所示。

全局模式

给定的匹配

预测新的匹配

图10-4 基于机器学习方法的全局模式与源模式映射

**2.** **适应性查询**

当一个提交给全局模式的查询重写为一系列的面向各个数据源的查询时，就需要有效 地执行这些查询。这时出现的问题是由于数据集成系统中的信息具有动态特性，导致数据 集成系统的数据源具有自治性和异构性，各个数据源数据的可访问性以及传输速度变化不 可预测，执行引擎没有足够的信息制订出一个标准的查询计划。针对这种情况，需要在查 询执行过程中动态调整查询计划，实现适应性查询处理。

**3.XML**

XML 是一种可扩展的标记语言，以开放的自描述方式定义数据结构，在描述数据内容 的同时能够突出数据结构描述，进而体现了数据之间的关系。XML 是一种半结构化数据模 型，可以用于描述不规则数据，集成来自不同数据源的数据，它也能够使来自不同数据源 的结构化数据容易地结合在一起。由于 XML 的存在，数据集成系统就不必了解每个数据 库描述数据的模式和规则，克服了 Web 数据源中的数据表现形式无穷尽问题。





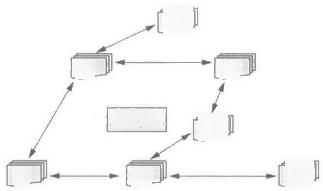
**第10章** **大数据预处理技术**

**4.P2P** **数据管理**

点对点 (Pee-to-Pee,P2P) 计算兴起和点对点文件共享系统的出现，促进了数据管理 领域应用 P2P 结构实现数据共享的研究，

P2P 计算提供了一种分布式管理共享数据的模式，每一个数据源仅需要提供自己和他 周围一系列邻近数据源的语义映射关系，其他更为复杂的集成是系统遵循网络中的语义路 径而形成，源的描述提供了研究P2P 结构下的模式及其映射建立的基础。

利用一个全局模式为数据集成系统服务是一个困难的问题。这是由于一个单独的全局 模式难以清楚地表述系统中全部的语义关系。但在P2P 结构下，没有一个单独的全局模式， 数据共享发生在网络上的数据源和邻近数据源之间。如图10-5所示是P2P 数据管理的实现。



Q3

Q4

os

06

LAY,GL

Q2

QI

Q

图10-5 P2P数据管理的实现

**10.7** **数据转换与约简**

在数据的预处理过程中，需要进行数据转换和数据约简处理，进而获得更高质量的数 据。这有助于后续的数据分析，并获得有价值的分析结果。

**10.7.1** **数据转换**

数据转换是指将原始数据变换成适合数据分析的形式。数据转换的主要内容有数据平 滑、数据聚集、数据泛化和数据规范化等。

**1.** **数据平滑**

数据平滑作用是指所获得的数据存在噪声不可避免，而通过数据平滑处理可以除掉数 据中存在的噪声以提高信噪比。数据平滑可以用分箱、聚类和回归方法来实现。

**2.** **聚集**

数据聚集就是汇总一批细节数据，归纳成一个更抽象的数据。

**3.** **数据泛化**

数据泛化过程即概念分层，将低层次的数据提炼到更高一级的概念层次中。





**大数据基础与应用**

**4.** **数据规范化**

数据规范化是将原来的度量值转换为无量纲的值。通过将属性数据按比例缩放，通过 一个函数将给定属性的整个值域映射到一个新的值域中，每个旧值都被一个新值替代。规 范化又分为最小—最大规范化、0值规范化和小数定标规范化。

最小一最大规范化保持原有数据之间的联系。如果输入值在原始数据值域之外，该方 法将为“越界错误”。0值规范化是指当某属性的实际最大值和最小值未知，或异常点左右 了最小一最大规范化时，该方法则有用。小数定标规范化是指通过移动属性的小数点位置 进行规范化。

**10.7.2** **数据约简**

数据约简是指在对挖掘任务和数据本身内容理解的基础之上，寻找依赖于发现目标的 数据的有用特征，以缩减数据规模，从而在尽可能保持数据原貌的前提下，最大限度地精 简数据量。数据约简技术可以用来得到数据集的约简表示，虽然数据集小，但仍接近于保 持原始数据的完整性。也就是说，在约简后的数据集上挖掘将更有效，仍然产生与约简前 相同(或几乎相同)的分析结果。

**1.** **数据约简的类型**

(1)特征约简

特征约简是在保留甚至提高原有判别能力的要求下，从原有的特征中删除不重要或不 相关的特征，或者通过对特征进行重组来减少特征的个数，同时减少特征向量的维度。特 征约简算法的输入是一组特征，输出是它的一个子集。在领域知识缺乏的情况下进行特征 约简的步骤如下：

① 搜索过程：在特征空间中搜索特征子集，每个子集称为一个状态，由选中的特 征构成。

② 评估过程：输入一个状态，通过评估函数或预先设定的阈值输出一个评估值，搜索 算法的目的是使评估值达到最优。

③ 分类过程：使用最终的特征集完成最后的计算。

(2)样本约简

如果已知样本数量很大，虽然样本质量参差不齐，但它们的确是实际问题的先验知识。 针对这种情况，就可以通过样本约简从数据集中选出一个有代表性的样本子集。子集大小 的确定要考虑计算成本、存储要求、估计量的精度以及其他与算法和数据特性有关的因素。

初始数据集中最大和最关键的维数就是样本的数目，也就是数据表中的记录数。数据 挖掘处理的初始数据集描述了一个极大的总体，对数据的分析只基于样本的一个子集。获 得数据的子集后，用它来提供整个数据集的信息，这个子集通常叫作估计量，它的质量依 赖于所选子集中的元素。取样过程存在取样误差，它不可避免。当子集的规模变大时，取 样误差一般将降低， 一个完整的数据集在理论上不存在取样误差。与针对整个数据集的数 据挖掘比较起来，样本约简具有减少成本、速度更快、范围更广的优点，有时甚至能获得





第10章 大数据预处理技术

更高的精度。

**2.** **数据约简的策略**

(1)数据立方体聚集：聚集操作用于数据立方体结构中的数据。

(2)属性子集选择：可以检测并删除不相关、弱相关或冗余的属性或维。 (3)维度约简：使用编码机制来减小数据集的规模。

(4)数值约简：用替代的、较小的数据替换数据或估计数据，例如参数模型(只需要 存放模型参数，而不是实际数据)或非参数方法，如聚类、抽样和使用直方图。

(5)离散化和概念分层产生：属性的原始数据值用区间值或较高层的概念替换。数据 离散化是一种数据约简形式，对于概念分层的自动产生有用。离散化和概念分层产生允许 挖掘多个抽象层的数据，是数据挖掘强有力的工具。

花费在数据约简上的计算时间不应超过或抵消在约简后的数据上挖掘所节省的时间。

**小结**

数据抽取、数据清洗和数据集成是大数据技术中的几个重要步骤，通过上述步骤处理 之后，可以获得较高的数据质量。对高质量的数据进行挖掘和分析，可以获得可靠的结果， 进而获得重要的价值。本章介绍了数据抽取、Web 数据抽取、数据质量清洗和数据集成技 术的内容，通过这些内容的学习，可以为系统学习大数据技术奠定重要基础。



**第11章** **大数据分析**

本章主要内容

大数据分析

数据分析概述

基本数据分析方法

高级数据分析方法

复合技术分析

大数据分析基础

大数据预测分析

大数据分析应用

|  |  |
| --- | --- |
| 数据分析的概念  数据分析的目的与意义  数据分析方法的分类  数据分析的类型  数据分析步骤  时间数列及动态分析法  相关分析  回归分析  判别分析  对应分析  预测分析  主成分分析  多维尺度分析  因子分析  方差分析 快速傅里叶变换  分类  聚类分析  预测学简介  预测原理  预测的步骤  预测技术分类 | 统计方法  指标对比分析法  分组分析法  综合评价分析法  指数分析法  平衡分析法  平滑和滤波  基线与峰值  可视化分析  数据挖掘  预测性分析  语义引擎  数据质量和数据管理  大数据的离线与在线分析 |

|  |  |
| --- | --- |
| 预测模型及分类  大数据预测分析要素  大数据预测分析的演化 大数据预测分析相关问题 舆情监测与分析 | 为客户提供服务 优化业务流程  改善生活  提高体育成绩  优化机器和设备性能 改善安全和执法 |

大数据分析平台与工具

大数据分析平台

大数据分析的工具

金融交易

电信业务

销售



**第11章大数据分析**

大数据技术是指可以从大数据中快速获得有价值信息的全部技术，大数据分析是大数 据技术中的最重要一环，只有分析数据才能获得更多智能的、深入的、有价值的信息。现 在越来越多的行业不断地产生大数据，而这些数据的属性(数量、数度、多样性和复杂性 等)呈现了不断增长的复杂性，因此，大数据的分析方法尤为重要，是决定最终信息是否 有价值的主要步骤。

**11.1** **数据分析概述**

计算机分析效率高，远远超过人脑的处理速度。对于海量数据，计算机能够处理，而 人类脑力处理不了。这是因为人的思维能力有限。因为人类视觉处理能力远远超过信息阅 读和逻辑思维能力，通过可视化办法可以将数据做成动态图像，使人脑能够处理更大规模 的数据。但是还有更大的数据，仅仅依靠人类视觉还是处理不过来，这就要依靠计算机去 处理，甚至要运用超级计算机处理。基于粒度的不同，采用不同的时间尺度，看到的信息 是不同的。需要进行全维度分析，这就要靠计算机辅助的人机综合系统，或者完全采用计 算机建模分析。这是科技发展带来的优势。当产生了抽象模型之后，就可以针对积累的数 据进行模式匹配和模式识别，从而进行针对某种特定需求的操作判断。例如一家销售化妆 品的公司，要寻找目标客户，就可以对积累的海量用户消费数据进行筛选，根据模型识别 和匹配出目标客户群，找到目标客户群之后，还需要进行有针对性的营销。由于不同用户 群经常接触的媒体类型不同，因此，寻找每个人的媒体接触习惯、接触时间，甚至能够预 测下一次将在什么地方接触到也不同，基于这些考虑，能够定向推动广告信息，就可以做 到有的放矢。

**11.1.1** **数据分析的概念**

数据分析是指用准确适宜的分析方法和工具来分析经过处理的数据，提取具有价值的 信息，进而形成有效的结论并通过可视化技术展现出来的过程。更具体地说，数据分析是 通过应用技术与工具来分析与理解数据，并应用一整套软件、系统和业务战略组成完整解 决方案，把自拥有的数据与用户产生的非结构化数据结合起来，总览全局。追溯每个业务 流程中产生的庞大数据，并进行共享整合分析，从而帮助决策者全面监控业务流程，预测 和塑造所预期的结果。

**11.1.2** **数据分析的目的与意义**

**1.** **数据分析的目的**

数据分析的目的是对杂乱无章的数据进行集中、萃取和提炼，进而找出所研究对象的 内在规律，发现其价值。



**大数据基础与应用**

**2.** **数据分析的意义**

真正有价值的信息就隐藏在看似繁乱的数据之中，但是要获得有用的数据，首先要进 行数据分析，获得对数据的认知。就像我们人类认识世界的方式一样，从实践获得的数据 中学习，逐渐就形成各种认识模型、各种逻辑关系。例如，判断一个用户的喜好、经济状 况、婚姻状况等，在认知的基础上就可以定向营销，从而达到营销效率的提升。

在产品的整个生命周期中，数据分析的过程是质量管理体系的支持过程，包括从市 场调研到售后服务和最终处置的各个过程都需要适当运用数据分析，以提升有效性。例 如 CEO 要通过市场调查，分析所得数据以判定市场动向，从而制订合适的生产及销售计 划。因此，数据分析有极广泛的应用范围。

**11.1.3** **数据分析方法的分类**

数据分析方法可以分为三类，即基本分析方法、高级分析方法和数据挖掘方法。基本 数据分析方法以基础的统计分析为主，主要包括对比分析、趋势分析、差异显著性检验、 分组分析法、结构分析、因素分析、交叉分析、综合评价分析和漏斗图分析等。高级分析 方法以建模理论为主，数据挖掘以数据仓库、机器学习等复合技术为主。

**11.1.4** **数据分析的类型**

数据分析主要分为探索性数据分析、定性数据分析、离线数据分析和在线数据分析等， 简述如下。

**1.** **探索性数据分析**

探索性数据分析是指为了形成值得假设的检验而对数据进行分析的一种方法，是对传 统统计学假设检验手段的补充。探索性数据分析侧重于在数据之中发现新的特征。

**2.** **定性数据分析**

定性数据分析是指定性资料分析、定性研究、照片、观察结果等非数值型数据(或者 说资料)的分析。

**3.** **离线数据分析**

离线数据分析是指将待分析的数据先存储于磁盘中，然后再进行数据分析，离线数据 分析用于较复杂和耗时的数据分析和批处理。

**4.** **在线数据分析**

在线数据分析 (OLAP, 也称为联机分析处理)用来处理用户的在线请求，它对响应时 间的要求比较高，通常处于秒级。与离线数据分析相比，在线数据分析能够实时处理用户 的请求，并且能够允许用户随时更改分析的约束和限制条件。尽管与离线数据分析相比， 在线数据分析能够处理的数据量要小得多，但随着技术的发展，当前的在线分析系统已经





**第11章** **大数据分析**

能够实时地处理数千万甚至数亿条记录。许多在线数据分析系统构建在以关系数据库为核 心的数据仓库之上。

**11.1.5** **数据分析步骤**

最初的数据可能杂乱无章而无规律，但通过作图、制表和用各种形式的拟合来计算某 些特征量，探索规律性的可能形式，这就需要研究用何种方式去寻找和揭示隐含在数据中 的规律性。在探索性分析的基础上提出几种模型，然后通过进一步的分析从中选择所需的 模型。通常使用数理统计方法对所选定模型或估计的可靠程度和精确程度做出推断。

数据分析的具体步骤主要包括识别信息需求、收集数据、分析数据、评价并改进数据 分析的有效性等活动，如下所述。

**1.** **识别信息需求**

识别信息需求可以为收集数据、分析数据提供清晰的目标，是确保数据分析过程有效 性的首要条件。

**2.** **收集数据**

有目的地收集数据是确保数据分析过程有效的基础。需要对收集数据的内容、渠道、 方法进行策划。主要考虑下述问题。

① 将识别信息需求转化为更具体的要求，例如评价供方时，需要收集的数据包括其过 程能力、测量系统不确定度等相关数据；

② 明确由谁在何时何处，通过何种渠道和方法收集数据；

③ 记录表应便于使用；

④ 采取有效措施，防止数据丢失和虚假数据对系统的干扰。

**3.** **分析数据**

分析数据是指将收集到的数据通过加工、整理和分析后，并将其转化为信息的过程， 常用的分析数据方法有排列图、因果图、分层法、调查表、散布图、直方图、控制图、关 联图、系统图、矩阵图、KJ 法、计划评审技术、PDPC 法、矩阵数据图。

**11.2** **基本数据分析方法**

基本数据分析方法是较经典的常用的方法，主要解决一般问题。

**11.2.1** **统计方法**

在自然科学中，统计学方法论是很重要的一个基础。 一旦把统计学和大数据融合在一 起，将颠覆传统的思维。统计学是收集、分析、表述和解释数据的科学，统计是人类对事





**大数据基础与应用**

物数量的认识而形成的定义，统计具有合计、总计之意。统计学指对某一现象的数据的搜 集、整理、计算、分析、解释和表述等活动。在实际应用中，统计一般包括统计工作、统 计资料和统计科学等内容。统计学的目标是从各种类型的数据中提取有价值的信息，给人 后见之明，与大数据的研究范围一致。

**1.** **统计工作**

统计工作指利用科学的方法搜集、整理和分析，提供关于某方面的数量资料工作的总 称，是统计的基础，也称统计实践，或统计活动。它是随着人类社会的发展和管理的需要 而产生和发展起来的。在现实生活中，统计工作作为一种认识社会经济现象总体和自然现 象总体的实践过程， 一般包括统计设计、统计调查、统计整理和统计分析四个环节。

**2.** **统计资料**

统计资料指通过统计工作取得的、用来反映现象的数据资料的总称。统计工作所取得 的各项数字资料及有关文字资料， 一般反映在统计表、统计图、统计手册、统计年鉴、统 计资料汇编和统计分析报告中，也称统计信息。其包括刚刚调查取得的原始资料和经过一 定程度整理、加工的次级资料。

**3.** **统计科学**

统计科学也称统计学，是统计工作经验的总结和理论概括，是系统化的知识体系。指 研究如何搜集、整理和分析统计资料的理论与方法。统计学是应用数学的一个分支，主要 通过利用概率论建立数学模型，收集所观察系统的数据，进行量化的分析、总结，并进行 推断和预测，为相关决策提供依据和参考。现已被广泛地应用在各门学科之中。

统计学又分为描述统计学和推断统计学。描述统计学是指给定一组数据，可以摘要并 且描述这份数据的统计学。推论统计学是指观察者以数据的形态建立出一个用以解释其随 机性和不确定性的数学模型，以之来推论研究中的步骤及母体。这两种用法都被称作应用 统计学。

上述三个方面内容联系紧密，统计资料是统计工作的成果，统计工作与统计科学之间 是实践与理论的关系。它们涉及的计算主要有均值、中位数、众数、正态分布、抽样、标 准差、概率论、检验、方差分析等。

**11.2.2** **指标对比分析法**

指标对比分析法又称比较分析法，是统计分析中最常用的方法，是通过有关的指标对 比来反映事物数量上差异和变化的方法。有比较才能鉴别。单独看一些指标，只能说明总 体的某些数量特征，得不出什么结论性的认识。经过比较，就可以对规模大小、水平高低、 速度快慢做出判断和评价。

指标对比分析分为静态比较和动态比较。静态比较是同一时间条件下对不同总体指标 的比较，如不同部门、不同地区、不同国家的比较，也叫横向比较。动态比较是同一总体 条件不同时期指标数值的比较，也叫纵向比较。这两种方法既可单独使用，也可结合使用。





**第11章** **大数据分析**

进行对比分析时，可以单独使用总量指标或相对指标或平均指标，也可将它们结合起来进 行对比。比较的结果可用相对数，如百分数、倍数、系数等，也可用相差的绝对数和相关 的百分点来表示。

**11.2.3** **分组分析法**

指标对比分析法是总体上的对比，但统计分析不仅要对总体数量特征和数量关系进行 分析，还要深入总体的内部进行分组分析。分组分析法就是根据统计分析的目的要求，把所 研究的总体按照一个或者几个标志划分为若干个部分，加以整理，进行观察、分析，以揭示 其内在的联系和规律性。统计分组法的关键问题在于正确选择分组标值和划分各组界限。

**11.2.4** **综合评价分析法**

社会经济现象往往错综复杂，社会经济运行状况是多种因素综合作用的结果，而且各 个因素的变动方向和变动程度是不同的。对宏观经济运行的评价，涉及生活、分配、流通、 消费各个方面；对企业经济效益的评价，涉及人、财、物合理利用和市场销售状况。如果 只用单一指标就难以得到恰当的评价，进行综合评价包括四个步骤。

① 确定评价指标体系，这是综合评价的基础和依据。要注意指标体系的全面性和系 统性。

② 搜集数据，并对不同计量单位的指标数值进行同度量处理。可采用相对化处理、函 数化处理、标准化处理等方法。

③ 确定各指标的权数，以保证评价的科学性。根据各个指标所处的地位和对总体影响 程度不同，需要对不同指标赋予不同的权重。

④ 对指标进行汇总，计算综合分值，并据此做出综合评价。

经济波动客观存在，任何国家都难以完全避免。如何避免大的经济波动，保持经济的 稳定发展， 一直是在宏观调控和决策中面临的重要课题，景气分析正是适应这一要求而产生 和发展的。景气分析是一种综合评价分析，可分为宏观经济景气分析和企业景气调查分析。

**11.2.5** **指数分析法**

指数是指反映社会经济现象变动情况的相对数。根据指数所研究的范围不同可以有个 体指数、类指数与总指数之分。指数综合反映复杂的社会经济现象的总体数量变动的方向 和程度，可以分析某种社会经济现象的总变动受各因素变动影响的程度。操作方法是通过 指数体系中的数量关系，假定其他因素不变，来观察某一因素的变动对总变动的影响。

**11.2.6** **平衡分析法**

平衡分析是研究数量变化对等关系的一种方法，它把对立统一的双方按其构成要素排 列起来，给人以整体的概念，以便于全局来观察它们之间的平衡关系。平衡关系广泛存在





**大数据基础与应用**

于经济生活中，大至宏观经济运行，小至个人经济收支。平衡种类繁多，如财政平衡表、 劳动力平衡表、能源平衡表、国际收支平衡表、投入产出平衡表等。平衡分析的作用如下。

① 从数量对等关系上反映现象的平衡状况，分析各种比例关系相适应状况。 ② 揭示不平衡的因素和发展潜力。

③ 利用平衡关系可以从各项已知指标中推算未知的个别指标。

**11.2.7** **平滑和滤波**

平滑和滤波是低频增强的空间域滤波技术。它的目的有两类： 一类是模糊；另一类是 消除噪声。空间域的平滑和滤波一般采用简单平均法进行，就是求邻近像点的平均亮度值。 邻域的大小与平滑的效果直接相关，邻域越大，平滑的效果越好，但邻域过大，平滑会使 边缘信息损失得越大，从而使输出的图像变得模糊，因此需合理选择邻域的大小。

**11.2.8** **基线与峰值**

**1.** **基线**

在计算机术语中，基线是项目存储库中每个工件版本在特定时期的一个快照。它提供 一个正式标准，随后的工作基于此标准，并且只有经过授权后才能变更这个标准。建立一 个初始基线后，以后每次对其进行的变更都将记录为一个差值，直到建成下一个基线。

**2.** **峰值**

峰值是指在所考虑的时间间隔内，变化量的最大瞬间值。峰值功率就是最高能支持的 功率。电源的峰值功率指电源短时间内能达到的最大功率，通常仅能维持30s 左右的时间。 一般情况下电源峰值功率可以超过最大输出功率50%左右，由于硬盘在启动状态下所需要 的能量远远大于其正常工作时的数值，因此系统经常利用这一缓冲为硬盘提供启动所需的 电流，启动到全速后就会恢复到正常水平。峰值功率其实没有什么实际意义，因为电源一 般不能在峰值输出时稳定工作。

**11.3** **高级数据分析方法**

**11.3.1** **时间数列及动态分析法**

将同一指标的一系列数值，按时间先后顺序排列，就形成时间数列，又称动态数列。 通过时间数列的编制和分析，可以找出动态变化规律，为预测未来的发展趋势提供依据。 时间数列可分为绝对数时间数列、相对数时间数列、平均数时间数列。

在统计分析中，如果只有单独的一个时期指标值，很难做出准确判断。如果编制了时 间数列，就可以进行动态分析，反映其发展水平和速度的变化规律及趋势。动态分析注意





**第11章大数据分析**

数列中各个指标具有的可比性。总体范围、指标计算方法、计算价格和计量单位，都应该 前后一致。时间间隔一般也要一致，但也可以根据研究目的，采取不同的间隔期，如按历 史时期分。为了消除时间间隔期不同而产生的指标数值不可比，可采用年平均数和年平均 发展速度来编制动态数列。

**11.3.2** **相关分析**

**1.** **相关分析的内容**

相关分析是一种研究变量相关性的统计方法，包括变量之间依存关系是否存在，存在 什么样的依存关系，以及相关程度和相关方向等。相关关系是一种非确定性的关系，例如， 以X 与 Y 分别记一个人的身高和体重，则X 与 Y 显然有关系，而又没有确切到可由其中 的一个精确地决定另一个的程度，这就是相关关系。在因果分析中有广泛应用，例如应用 相关分析判断经济指标之间的替代关系和关联度。相关分析可以用来研究两个变量的关系， 测定它们之间联系的紧密程度。

**2.** **相关分析的分类**

(1)线性相关分析

如果两个变量变化的方向一致，则称为正相关；如果两个变量变化的方向相反，则称 为负相关；否则为无线性相关。

(2)偏相关分析

在控制一些对两变量之间的相关性可能有关的其他变量之后，再对两变量的线性相关 性分析。例如通过控制年龄和学术水平的影响，估计工资收入和受教育水平的相互关系。

(3)距离分析

可以通过距离的大小对两变量之间相似或不相似程度的测度，在这里所提及的距离可 以是观测量之间的距离和变量之间的距离，例如欧式距离和海明距离等。

**11.3.3** **回归分析**

回归分析是在掌握大量观察数据的基础之上，利用数理统计方法建立变量与自变量之 间的回归关系函数表达式，即回归方程式。

回归分析是研究一个随机变量Y 对另一个X 变量或一组变量 (Xj,X₂,…,Xr) 的相依 关系的统计分析方法，即回归分析是确定两种或两种以上变数间相互依赖的定量关系的一 种统计分析方法。回归分析主要用于得到变量之间的关系，即变量是否相关联、相关的方 向和相关的强度等。之后建立响应的数学模型，对感兴趣的变量预测，找出能够代表所有 观测资料的函数曲线，然后用此函数表示变量与自变量之间的关系。

相关分析与回归分析虽然都是分析变量之间的相互关系，但是相关分析是回归分析的 基础，回归分析是认识变量之间相互程度的具体形式。回归分析的步骤如下：

① 确定自变量与因变量。



**大数据基础与应用**

② 根据自变量与因变量的历史统计资料进行计算，建立回归分析预测模型。 ③ 获得自变量与因变量之间的某种因果关系。

④ 模型检验，预测误差是衡量一个预测模型的重要指标，小误差表明模型可以得到比 较好的预测结果。

⑤ 运用之前确定的回归预测模型进行预测计算，再根据具体的实际情况，运用相关知 识进行全面分析，进而得到最终的预测值。

回归分析按照涉及的自变量是一个或多个，可分为一元回归分析和多元回归分析。按 照自变量和因变量之间的关系类型，又可分为线性回归分析和非线性回归分析。

**11.3.4 判别分析**

判别分析又称为分辨法，其分类方式是事先确定的，根据若干变量值判断研究对象归 属问题的一种多变量统计分析方法。判别分析的原理是按照一定的判别原则建立一个或多 个判别函数，通过研究与对象的大量相关资料确定函数中的变量系数，并计算判别指标， 通过判别指标确定观察对象的数列。其主要目的是根据一致类别的样本做出判别模型，然 后根据模型来判别未知类别的样本的归属。

**11.3.5** **对应分析**

对应分析又称为关联分析，是一种多元统计分析技术，对有定性变量构成的交互汇总 表进行分析，进而发现变量之间的联系。这种分析主要适于多类别的类变量，可以发现同 一变量的各类别之间的差异，以及不同变量各类别之间的对应关系。对应分析的原理是将 一个列联表的行和列中各元素的比例结构以点的形式在较低维的空间中表示出来。其最大 特点是把大量样本和变量同时展现在一张图上，将样本的大类及其属性在图上直观表现出 来，这种图示化技术是市场分析技术的强有力工具。

**11.3.6** **预测分析**

根据已知的过去和现在推测未来就是预测分析。统计预测属于定量预测，是以数据分 析为主，在预测中结合定性分析。统计预测的方法大致可分为两类： 一类是主要根据指标 时间数列自身变化与时间的依存关系进行预测，属于时间数列分析；另一类是根据指标之 间相互影响的因果关系进行预测，属于回归分析。

预测分析的方法有回归分析法、滑动平均法、指数平滑法、周期变化分析和随机变化 分析等。比较复杂的预测分析需要建立计量模型，求解模型中的参数有多种方法。

**11.3.7** **主成分分析**

主成分分析是一种多元统计方法，其基本思想是将多个变量进行线性变换，进而得到 不相关的综合变量，然后根据给定的规则从中选择出少数几个能够较好反映出原始变量信





第 1 1 章 大数据分析

息的综合变量。主成分分析是通过对协方差矩阵进行特征分解，以得出数据的主要成分， 即特征向量及其特征值。主成分分析的基本步骤如下：

① 计算相关系数矩阵；

② 求出相关系数矩阵的特征值及相应的正交化单位特征向量；

③ 选择主成分；

④ 计算主成分得分。

**11.3.8** **多维尺度分析**

多维尺度分析将多维空间的研究对象简化到低维空间进行定位、分析和归类，同时又 保留对象间的原始关系的数据分析方法。例如对于消费者行为分析问题，可以将消费者对 品牌的感觉偏好以点的形式在多维空间上表示，对不同品牌感觉的差异是用点的距离体现。 将这种品牌或项目的空间定位点图称为空间图，空间轴表示消费者形成对品牌的感觉或偏 好的各种因素或变量。多维尺度分析方法的基本思想是被访问者对研究对象的感知，是以 被访问者对研究对象的分组来反映的。多维尺度分析是分析消费者和偏好的最有效的方法， 具有直观性和合理性。

**11.3.9** **因子分析**

因子分析是指研究从变量群中提取共性因子的统计技术。因子分析就是从大量的数据 中寻找内在的联系，减少决策的困难。因子分析的方法有十多种，如重心法、影像分析法、 最大似然解、最小平方法、阿尔发抽因法、拉奥典型抽因法等。这些方法本质上都属于近 似方法，以相关系数矩阵为基础。

**11.3.10** **方差分析**

方差分析又称为变异数分析，主要用于两个及两个以上样本均数差别的显著性检验。 由于各种因素的影响，研究所得的数据呈波动状。造成波动的原因可分成两类， 一类是不 可控的随机因素，另一类是研究中施加的、能对结果形成影响的可控因素。方差分析是从 观测变量的方差入手，研究各控制变量中对观测变量有较显著影响的变量。

**11.4 复合技术分析**

统计方法往往有能力的极限，如只用统计机器翻译方法，翻译质量的提高就有限度。 一种可能的途径是把其他方法和统计方法结合起来，采用多元化的方法来建立综合性模 型。传统 AI (如机器学习)先通过在较小的数据样本集学习，验证分类、判定等“假设” 和“模型”的适合性，再应用推广到更大的数据集，而大数据机器学习选择大样本和全 样本技术。



**大数据基础与应用**

**11.4.1** **快速傅里叶变换**

有限长序列可以通过离散傅里叶变换 (DFT) 将其频域也离散化成有限长序列，但其 计算量太大，很难实时地处理问题，因此引出了快速傅里叶变换 (FFT)。1965 年 ，Cooley 和Tukey提出了计算离散傅里叶变换 (DFT) 的快速算法，将 DFT 的运算量减少了几个数 量级。从此，对快速傅里叶变换 (FFT) 算法的研究便不断深入，数字信号处理这门新兴 学科也随 FFT 的出现和发展而迅速发展。根据对序列分解与选取方法的不同而产生了FFT 的多种算法，基本算法是基于DFT 。FFT 在离散傅里叶反变换、线性卷积和线性相关等方 面也有重要应用。

快速傅氏变换 (FFT) 是离散傅氏变换的快速算法，它是根据离散傅氏变换的奇、偶、 虚、实等特性，对离散傅立叶变换的算法进行改进而得。它对傅氏变换的理论并没有新的 发现，但是在计算机系统或者数字系统中应用离散傅立叶变换，却前进了一大步。

**11.4.2** **分类**

分类就是找出一个类别的概念描述，它代表了某类数据的内涵描述，并用这种描述来 构造模型， 一般用规则或决策树模式表示。分类是利用训练数据集通过一定的算法而求得 分类规则，分类可用规则描述和预测。分类过程是：首先从数据中选出已经分好类的训练 集，在该训练集上运用数据挖掘分类的技术，建立分类模型，对于没有分类的数据进行分 类。类的个数是确定的，预先定义好的。例如：

① 信用卡申请者按低、中、高风险分类。

② 对钢材生产的全流程进行质量监控和分析，构建故障地图，实时分析产品出现瑕疵 的原因分类，有效提高了产品的优良率。

**11.4.3** **聚类分析**

聚类分析指将物理或抽象对象的集合分组成为由类似的对象组成的多个类的分析过 程。聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大 的相似性，而不同簇间的对象有很大的相异性。聚类分析是一种探索性的分析，在聚类的 过程中，人们不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分 类。聚类分析所使用的方法不同，常会得到不同的结论。不同研究者对于同一组数据进行 聚类分析，所得到的聚类数未必一致。在商业中，聚类可以帮助市场分析人员从消费者数 据库中区分出不同的消费群体，并且概括出每一类消费者的消费模式或者消费习惯。它作 为数据挖掘中的一个模块，可以作为一个单独的工具以发现数据库中分布的一些深层次的 信息，或者把注意力放在某一个特定的类上以做进一步的分析并概括出每一类数据的特点。 图11-1是聚类算法的一种展示，在二维空间上，根据距离的远近，可将对象聚为3类。





第 1 1 章 大数据分析

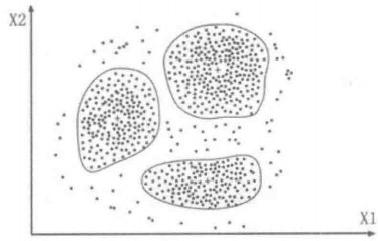


图11-1 聚类算法示意图

**11.5** **大数据分析基础**

与传统的数据统计分析相比较，大数据的5V 特点，使得大数据分析必然要依托计算 机科学技术来实现，进而逐渐出现了大数据分析的两个研究方向。 一个方向侧重于数据处 理与表示，主要强调采集、存储、处理和数据可视化方法的研究。另一方向侧重于数据统 计规律，注重对微观数据本质特征的提取与模式发现。现在的大数据分析逐渐由数据处理 技术向数据分析的倾斜。

只有通过对大数据分析才能获取很多智能的、深入的、有价值的信息。越来越多的应 用涉及大数据，而这些大数据的属性，包括数量、速度、多样性等都是呈现了大数据不断 增长的复杂性，所以大数据的分析方法在大数据领域就显得尤为重要，是决定最终信息是 否有价值的决定性因素。大数据分析普遍存在的方法理论有可视化分析、数据挖掘、预测 性分析、语义引擎、数据质量和数据管理。

**11.5.1** **可视化分析**

数据分析结果的使用者有大数据分析专家，同时还有普通用户，但是他们二者对于大 数据分析最基本的要求就是可视化，因为可视化能够直观地呈现大数据特点，同时能够非 常容易被读者所接受，就如同看图说话一样简单明了。不管是对数据分析专家还是普通用 户，数据可视化是数据分析工具最基本的要求。可视化可以直观地展示数据，让数据自己 说明，让观众看到结果。

**11.5.2** **数据挖掘**

大数据分析的理论核心就是数据挖掘，各种数据挖掘的算法基于不同的数据类型和格 式才能更加科学地呈现出数据本身具备的特点，也正是因为这些被统计学家所公认的各种 统计方法才能深入数据内部，挖掘出公认的价值。也是因为有这些数据挖掘的算法才能更 快速地处理大数据。如果一个算法得用上好几年才能得出结论，那大数据的价值也就无从





**大数据基础与应用**

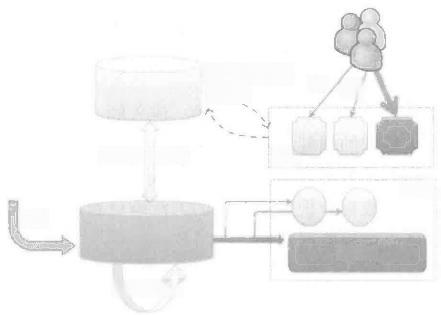
说起了。可视化是给人看的，数据挖掘就是给机器看的。集群、分割、孤立点分析还有其 他的算法可以使我们深入数据内部，挖掘价值。这些算法不仅能够处理大数据的数据量， 也能获得处理大数据的理想速度。

**11.5.3** **预测性分析**

数据挖掘能使分析员更好地理解数据，而预测性分析可以让分析员根据可视化分析和 数据挖掘的结果做出一些预测性的判断。

**11.5.4** **语义引擎**

由于非结构化数据与异构数据等的多样性带来了数据分析的新的挑战与困难，需要一 系列的工具去解析、提取、分析数据。语义引擎需要被设计成能够从文档中智能提取信息。 从大数据中挖掘出特点，通过科学的建立模型，之后便可以通过模型输入新的数据，从而 预测未来的数据。图11-2描述了搜索引擎的迭代过程。



Online Lcarning

A/Btst

策略

数据

网页

网页库

倒排表

Data Mining

Feature Training

机器学习平台

s%

5%

图11-2 搜索引擎的迭代过程

**11.5.5** **数据质量和数据管理**

大数据分析离不开数据质量和数据管理，高质量的数据和有效的数据管理，都能够保证分 析结果的真实和价值。

上述几点是大数据分析基础，针对具体的需要，将构建更加有特点的、更加专业的大 数据分析方法。

**11.5.6** **大数据的离线与在线分析**

**1.** **大数据的离线分析**

由于大数据的数据规模已经远超出当今单个计算机的存储和处理能力，当前的离线数





第 1 1 章 大 数 据 分 析

据分析通常构建在云计算平台之上，并且应用离线分析，如开源的 Hadoop 的 HDFS 文件 系统和 MapReduce 运算框架。Hadoop 机群包含数百台乃至数千台服务器，存储了数PB 乃 至数十倍的 PB 数据，每天运行着成千上万的离线数据分析作业，即 MapReduce 任务。每 个作业处理几百 MB 到几百 TB 的数据甚至更多的数据，作业运行时间从几分钟到几个小 时、几天甚至更长的时间。

大数据的离线分析是将大数据存入分布系统的磁盘，然后再进行分析处理。例如批量 处理就是典型的离线分析。而流式大数据分析就不是离线分析，而是在线分析。

**2.** **大数据的在线分析**

当前的在线分析系统已经能够实时地处理数千万甚至数亿条记录。许多在线数据分析 系统构建在以关系数据库为核心的数据仓库之上，不少新兴的在线数据分析系统构建在云 计算平台之上的NoSQL 系统，例如 Hadoop 上的 HBase。因为后者开源、易于扩展和管理， 而且成本更低。如果没有大数据的分析和处理，则无法存储和索引数量庞大的互联网网页， 就不会有当今的搜索引擎。大数据的分析处理还在很多其他地方发挥着作用，与许多市场 调研公司通过电话、邮件、信函等方式进行抽样调查和分析不同的是，这些数据都是通过 对实际访问、交易的大量真实数据进行统计分析得到的。

**11.6** **大数据预测分析**

大数据技术的本质是能够解决实际问题，预测是大数据分析技术的最重要应用之一。 大数据的核心价值就是预测，通过预测可以获得正确的判断，进而创造巨大的价值。大数 据预测分析是有目的地收集数据、分析数据，进而获取信息的过程，也就是为了解决生活 与生产中的决策等问题，运用分析方法对数据进行处理的过程。数据统计分析就是运用统 计学的方法对数据进行处理，数据统计分析能够挖掘出数据中隐藏的信息，但是这种数据 的分析是向前分析，是分析已经发生过的事情。而在大数据中，数据统计分析是向后分析， 因此，大数据的统计分析具有预见性。

预测分析是根据可视化分析和数据挖掘的结果做出预测性的判断。可视化分析和数据 挖掘是预测分析的基础，只要在大数据中挖掘出信息的特点与联系，就可以建立科学的数 据模型，通过模型输入新的数据，从而预测未来的数据。

**11.6.1** **预测学简介**

预测分析是指在调查研究的基础上，研究某事物的发展变化的规律，进而对事物的未 来进行科学的预判与分析。在预测分析中所使用的方法和手段统称为预测技术。预测分析 是指预测理论。将预测理论和预测技术集成为预测科学，简称为预测学。

预测学的目的是减少由于不确定性导致错误决策所产生的风险。预测理论致力于对不 知的和随机的后果进行数学分析和描述，预测技术运用模型，为决策者提供恰当决策的必





**大数据基础与应用**

要信息。不确定性是所有问题的产生的根源。为将非现实的、确定的世界观转变为现实、 随机的世界观铺平了道路。

传统预测学是将阴阳、五行、周易、八卦、奇门遁甲等集成，以推测已知或未知的事 件为目的的一门学科。现代预测学主要研究无所不在的不确定性，目的是控制随机性以及 减少无知的程度。预测学通过开发数学模型和程序，制定事物未来发展的可靠预测，揭示 过去发生事件的准确结果。

**1.** **科学解析**

客观事物的可预测性是预测科学的基本前提。事物的出现都由一定的原因产生。也就 是说，产生事物的原因在该事物产生之前就必然存在，否则该事物就决不能产生。因此， 事物产生之前只要了解了产生该事物的全部的或主要的原因，或者这些原因的原因等，就 可以预知该事物的产生。基于这一点，客观事物具有可预测性，这是包括现代天气预测、 地震灾害预测在内的一切科学预测的最基本的前提。预测是一个科学的命题，是辩证唯物 主义可知论的具体应用。

**2.** **古典奇门遁甲预测的方法论**

奇门遁甲预测是中国古代术数理论，被称为《易经》最高层次的预测学，其本质是一 门高等的天文物理学，揭示了太阳系八大行星和地球磁场的作用情况。奇门就是指八门， 遁甲就是甲这个天干在奇门局里隐藏不见。奇门遁甲的预测准确度很高，有古语为证：学 会奇门遁，来人不用问。

(1)古典奇门遁甲预测的方法论的内容

① 根据客观事物的不同属性，将其进行分类，然后用符号表示各类事物。 ② 利用符号之间的关系来代换其所代表的具体事物之间的关系。

③ 运用五行生克规律进行推导，并把符号之间的推导结论还原为具体事物发展的 结论。

在现实生活中，在某一事物产生之前，要知道产生该事物全部的或主要的原因十分困 难，这就为预测提出了一个问题。这也正是利用现代数学方法，如概率论等进行预测时的 最大障碍。古典奇门遁甲预测的方法论，实质上是一个以客观事物可预测性为基本前提的 符号预测方法体系。这种方法的核心是具有一定属性(阴阳五行属性)的符号对具有相应 属性事物的代换。在这一代换体系中，不同事物与时间都被代换成相应的符号，都具有相 应的属性。不同的时间横断面的符号体系与各种不同事物之间建立起了对应关系。于是， 人们就可通过观察不同时间横断面的符号之间的关系，运用五行生克规律推导出一定的结 论，并把它还原为具体事物发展的结论。

(2)奇门遁甲预测的方法论的讨论

五行生克规律在中医学领域的成功应用(阴阳五行学术是中医学的主要理论),说明五 行生克所用的推导规律，基本上能反映客观事物的运动规律。

在不进行充分实证的情况下，符号与具体事物之间的代表与被代表关系，只能是一些 没有经过科学证明的假说。





第11章 大数据分析

假说在未被证明之前，则有三种可能的情况。

① 假说全部成立；

② 假说全部不成立；

③ 假说一部分不成立， 一部分成立。

显然，第一种情况不可能；第二种情况也不可轻下结论。因为如果现在就做出结论， 确认第二种情况的话，则这一结论就是未经任何实践检验和科学证明的结论。而不经实践 检验和科学证明的结论是不科学的结论。所以，只有第三种情况才符合科学的实验精神和 科学的探索精神。而从第三种情况出发，就可以做出这样的科学猜想，古典奇门遁甲预测 的方法论，也许包含有一定的科学成分，值得研究探讨。

上述关于古典奇门遁甲预测的方法论的认识，也是对于所有假说的认识，具有广泛性 和普适性。例如六度分隔假说、CAP 理论，在没有科学证明之前也是假说。

**3.** **预测技术应用**

预测决策理论和方法渐渐被引入到了工业安全领域，用以科学指导安全生产，并取得 了一定成效。特别是目前随着现代数学方法和计算机技术的发展，国际上安全评价分析以 及预测决策实施得到了广泛应用，如模糊故障树分析预测、模糊概率分析、模糊灰色预测 决策等。利用计算机专家系统、决策支持系统、人工神经网络等现代数学方法和计算机技 术，使安全分析评价预测决策开拓了一个更广阔的应用前景，这些技术方法在核工业、化 工、环境等领域得到了广泛应用。以安全分析、隐患评价、事故预测决策为主体的安全评 价工作作为一种产业在国际上已经出现。现在，大数据技术的出现，为预测展现了可能性， 人们对其寄予无限的希望。

**11.6.2** **预测原理**

预测原理主要包括整体性原理、可知性原理、可能性原理、相似性原理和反馈原理等。 (1)整体性原理

事物是由若干部分相互关联而成的有机整体，事物发展变化的过程也是一个有机整体 发展变化的过程，因此以整体性为特征的系统理论是预测的基本理论。

(2)可知性原理

由于事物发展过程的统一性，即事物发展的过去、现在和将来是一个统一的整体，所 以人类不但可以认识预测对象的过去和现在，而且也可以通过过去到现在的发展规律，推 测将来的发展变化。

(3)可能性原理

预测对象的发展有各种各样的可能性，预测是对预测对象发展的各种可能性的一种估 计。如果认为预测是必然结果，则失去了预测的意义。

(4)相似性原理

预测对象与已知事物的发展变化规律相类似，把预测对象与类似的已知事物的发展变 化规律进行类比，就可以对预测对象进行描述。





**大数据基础与应用**

(5)反馈原理

预测未来的目的是为了更好地指导当前，因此应用反馈原理不断地修正预测才会更好 地指导当前工作，为决策提供依据。

**11.6.3** **预测的步骤**

科学的预测是在广泛调查研究的基础上进行的，涉及方法的选择、资料的收集、数据 的整理、建立预测模型、利用模型预测和对预测结果进行分析等一系列工作。预测的基本 步骤归纳如下。

**1.** **确定目标**

在该阶段，确定预测对象、预测目的和目标，明确预测要求等。

**2.** **选择预测方法**

预测方法很多，到目前为止，各类预测方法不下几百种。因此应根据预测的目的和要 求，考虑预测工作的组织情况，合理地选择效果较好的、既经济又方便的一种或几种预测 方法。

**3.** **收集和分析数据**

该阶段根据预测目标和所选择预测方法的要求去收集所需原始数据。原始数据是进行预 测的重要依据，所收集原始数据的质量和可靠性将直接影响预测的结果。对原始数据的要求 是数据量足够、质量高，只有这样，才能较准确地反映事物的规律。因此收集足够数量的可 靠性高的数据是这一阶段的任务。数据的分析和整理是发现系统发展变化规律性和系统各组 成部分内在联系的关键，是建立预测模型的根据，因此要选择合适的数据处理方法。

**4.** **建立预测模型**

建立预测模型是预测的关键工作，它取决于所选择的预测方法和所收集到的数据。建 立模型的过程分为建立模型和模型的检验分析两个阶段。在建立预测模型时，只建模型， 不进行检验，这样的预测是不令人信服的。只有通过检验的模型，才能用于预测。

**5.** **模型的分析**

模型的分析是指对系统内部、外部的因素进行评定，找出使系统转变的内部因素和客 观环境对系统的影响，以分析预测对象的整体规律性和模型的适用性。

**6.** **利用模型预测**

所建立的模型是在一定假设条件下得到的，因此也只适用于一定条件和一定预测期限。 如果将其推广到更大范围，就要利用分析、类比、推理等方法来确定模型的适用性。只有 在确认模型符合预测要求时，才可利用模型进行预测和模型的适用性。





**第11章** **大数据分析**

**7.** **预测结果的分析**

利用预测模型所得的预测结果并不一定与实际情况符合。因为在建立模型时，往往有 些因素考虑不周，因资料缺乏以及在处理系统问题时的片面性等，使预测结果与实际情况 偏离较大，所以需要从下述两个方面进行分析。

① 用多种预测方法预测同一事物，将预测结果进行对比分析、综合研究之后加以修正 和改进；

② 应用反馈原理及时用实际数据修正模型，使预测模型更完善。

**11.6.4** **预测技术分类**

随着预测科学的发展，预测科学的应用领域越来越广，预测方法也越来越多。据估计， 目前预测方法已达150多种，其中较常用的有十几种。基于不同的分类原则对预测技术的 分类如下。

按预测技术的属性分类，预测技术可分为定性预测技术、定量预测技术和定时预测 技术。

按预测对象分类，预测技术可分为科学预测、社会预测、经济预测和市场预测等。

按预测方式分类，预测技术可分为直观性预测、探索性预测、目标预测和反馈预测等。

按预测时间分类，预测技术可分为短期预测、中期预测、长期预测和未来预测。 下面介绍几种预测技术。

**1.** **定性预测技术**

在没有较充分的数据可利用时，只能凭借直观材料，依靠个人经验和分析能力，进行 逻辑判断，对未来做出预测，即为定性判断预测技术。定性预测技术是以预测者的经验为 基础，判断发展趋势、探讨发展变化规律的方法，它适用于缺乏数据的情况下对事物的预 测。定性预测的优点是方法简便、灵活。实践中，有时即使有充足的数量，也采用定性预 测技术，其原因是把定性预测的结论与定量预测的结果相比较可以提高预测的准确性，同 时在定性预测的指导下进行定量预测可起到定量预测起不到的作用。常用的定性预测技术 有德尔菲法、主观概率法、交叉概率法、专家意见法等。

(1)德尔菲法

德尔菲法是美国兰德公司研究人员赫尔马 (O.Helmet) 和达尔奇 (N.Dalkey) 于 2 0 世纪40年代开发的一种预测方法。目前该方法已广泛用于军事预测、人口预测、医疗卫生 保健预测、经营和需求预测、教育预测以及方案评价的决策分析等领域。该方法的过程是 预测机构或人员预先选定与预测问题有关的专家10～15人，采用信件往来的方式与其建立 联系，将他们的意见进行整理、综合、归纳后再匿名反馈给各位专家再次征求意见，按这 种方式多次反复，直至使专家们的意见趋于一致为止，最后得出预测结论。该方法的具体 步骤如下。

① 选择专家。专家人数的确定依据所预测问题的复杂性和所需知识面的宽窄， 一般以 10～15人为宜。所选择的专家彼此不发生联系，只用书信的方式与预测人员直接发生联系。



**大数据基础与应用**

② 编制并邮寄专家应答表。首次交往需向专家介绍预测的目的，提供现有的相关资料， 并邮寄“专家应答表”。为避免浪费专家的时间，“专家应答表”应力求简练，只需专家用 “是”“否”等简单词句或符号回答或给予简单的评分。

③ 分析整理。“专家应答表”收集专家的意见，整理“专家应答表”,进行综合、分析、 归纳等工作。

④ 与专家反复交换意见。将整理、分析、归纳和综合的结果反馈给各专家并进一步提 供有关资料，专家修订自己的意见，再填写“专家应答表”,如此反复进行，直至得出预测 结论。

⑤ 将最终预测结论函告各专家并致谢。

采用德尔菲法的好处：

● 可以消除召开专家讨论会所出现的随声附和、崇拜专家、固持己见和有顾虑等 弊病。

● 可使意见迅速集中。

这种方法是在假设预测项目的各因素之间无交互作用的前提下进行的，因此有一定的 局限性。

(2)主观概率法

主观概率是某人对某事件发生可能性的主观估计值。对同一事物来说，不同的人因知 识、阅历、看问题的角度的不同等原因导致对问题的估计值也不同，这就是主观因素在起 作用。主观概率法就是在调查专家主观概率的基础上，寻求最佳主观估计的科学方法。如 果要预测某一事件发生的可能性，先调查一组专家的主观概率，然后加权平均即得某事件 发生的概率。

(3)交叉概率法

交叉概率法是对交互影响因素作用下的事物进行预测的一种定性预测技术。某事物的 发生或发展对其他事物将产生各种各样的影响，根据各事物之间的相互影响确定事物发生 的概率，并用以修正专家的主观概率，从而对事物的发展做出较客观的评价。这是该方法 的基本思想。该方法的步骤如下。

① 确定各事物之间的影响关系。

② 确定各事物之间的影响程度。

③ 计算某事物发生时对其他事物发生概率的影响。

④ 分析其他事件对该事件的影响。

⑤ 确定修正后的主观概率。

**2.** **定量预测技术**

由于被研究的系统复杂，只靠经验对其进行定性预测不够，还必须从数量上研究系统 的变化，因此还需研究定量预测技术。

数量分析是利用统计资料，借助数学工具，分析因果关系进行预测。数量分析预测方 法较多，如趋向外推法和回归分析法等。趋向外推法即时间序列分析法，它是根据历史和 现存的资料推测发展趋势，从而分析出事物未来的发展情况。它把在一定条件下出现的事





**第11章** **大数据分析**

件按时间顺序加以排列，通过趋势外推的数学模型预测未来。时间序列就是把统计资料按 发生的时间先后进行排列所得到的一连串数字。时序分析是研究预测目标与时间过程之间 的演变关系。因此它是一种定时的预测技术。回归分析法是从事物变化的因果关系出发来 进行预测。回归分析也称相关分析，是研究引起未来变化的各种客观因素的相互作用、指 出各种客观因素与未来状态之间统计关系的方法。

**11.6.5** **预测模型及分类**

预测模型是预测的核心，建立预测模型是预测技术的核心技术。

按变量之间的关系分类，预测模型可分为因果关系模型、时间关系模型和结构关系模 型等。

按变量形式分类，预测模型可分为线性预测模型和非线性模型。

按变量的数量分类，预测模型可分为一元模型和多元模型等。

按变量的性质分类，预测模型可分为定量因子模型、含定性因子模型和定性因子模 型等。

**11.6.6** **大数据预测分析要素**

实施成功的预测分析有赖于以下关键因素。

**1.** **数据质量**

数据是预测分析的血液。通常是来自内部的数据，如客户交易数据和生产数据。但还 需要补充外部数据源，如行业市场数据、社交网络数据和其他统计数据。与流行的技术观 点不同，这些外部数据未必一定是大数据。数据中的变量是否有助于有效预测才是关键所 在。总之，数据越多，相关度和质量越高，找出原因和结果的可能性越大。

**2.** **数据分析师**

数据分析师必须理解业务需求和业务目标，审视数据，并围绕业务目标建立预测分析 规则，例如增加电子商务的销售额、保持生产线的正常运转、防止库存短缺等。数据分析 师需要拥有数学、统计学等多个领域的知识。

**3.** **预测分析软件**

预测分析必须借助预测分析软件来评估他们的分析模型和规则，预测分析软件通过整 合统计分析和机器学习算法发挥作用。IBM SPSS和 SAS 是两个常用的分析软件。R 项目 则是一个非常流行的开源工具。如果数据量达到大数据的规模，那么就需要一些专门的大 数据处理平台，如Hadoop 或数据库分析机、Oracle Exadata 等。

**4.** **运营软件**

如果找到或建立了合适的预测规则，下一步就是将规则植入应用。预测分析软件应该





**大数据基础与应用**

能以某种方式产生代码，例如预测分析某产品。更重要的是将预测规则需要的数据事先准 备好。预测规则也能通过业务规则管理系统和复杂事件处理平台进行优化。

**11.6.7** **大数据预测分析的演化**

数据分析的传统范型 (BI-ETL-EDW) 即将被新的分析范型取代，这已达成共识。全 新的数据分析平台将消除当前分析软件在设计和实施方面的延迟和低效率，新的数据分析 范型是目标导向的，不关心数据的来源和格式，能够无缝处理结构化、非结构化和半结构 化数据。它能够输出有效结果，能够提供去黑箱化的预测分析服务，能够面向更广泛的普 通员工快速部署分析应用。数据管理、分析透明度以及用户应用是阻碍企业数据分析应用 的关键问题。

**1.** **数据管理**

Hadoop 已经成为管理大数据的基础支撑技术，并传递出一个非常明确的信号： Hadoop 发行商想要在Hadoop HDFS 之上提供实时、互动的查询服务。这个趋势将 SQL查询处理与 具备指数级扩展能力的 HDFS 存储架构整合到了一起，实现对PB 级别大数据的管理方式。

**2.** **去黑箱化**

预测和统计分析领域已经出现很多预测技术。但预测分析面临的最大问题是黑箱化。 随着企业领导越来越多地应用预测分析技术做出重大商业决策，预测分析技术需要去黑箱 化，有利于企业管理者彻底驾驭数据分析工具。不但看到数据分析结果，而且知道分析是 如何得来的、分析工具的设计原理等，这有助于管理者增加对预测分析的信心，而不是过 去那样完全不明就里。R 和 Stata 的崛起，正在冲击传统的黑箱式分析方法，这也代表着商 业世界的发展趋势。

**3.** **应用普及**

即使实现了分析的去黑箱化，企业数据分析应用在企业中的部署依然面临以下几个方 面的挑战：发布可复用应用、创建最佳实践、组织范围内的横向协作、无缝重组模型等。 在最终用户的应用普及是数据分析成功的关键。例如建设一个专门提供分析应用的企业移 动应用商店可能较显著地加快数据分析的应用普及。数据分析应用将不再仅是数据科学家 的工作，更多分析应用将以预先打包内容和应用发送到分析工程师手中。

**11.6.8** **大数据预测分析相关问题**

利用大数据较常用的工具，可以使用非交易数据来做出战略性的长期的业务决定。客 户服务代表可以独立决定一个问题客户是否值得保留或者升级，或者销售人员可以基于人 们对零售商在网站上的评价来调整零售商的产品量。大数据并不是要取代传统商务智能工 具，大数据将使BI 更有价值和更有利于业务发展。在大数据时代，因为数据很多，很可能 可以找到相关的关系，使原来找不到的相关关系现在找得到了。但是因为数据太多，不一





**第11章** **大数据分析**

定能够理解为什么。只要能先找到相关关系，就已经很不错了，也可能会找到内在的因果 关系。在大数据时代，不一定知其所以然，但知其然。

在预测中，如果仔细地查看使用从BI 工具收集到的历史交易数据，就会发现最新商家 定位活动更倾向于参考来自大数据技术处理的结果。

**1.** **分析社交媒体中的非结构数据**

在社交媒体中存在很大的商机，如果没有预测分析，很可能错过机会。在过去，根据 历史数据来做决定，但现在需要预测分析。需要结合大数据开源技术(大多数大数据平台 都源自开源)、摩尔定律、商品硬件、云计算以及捕捉和存储大量非交易数据来实现预测目 的。要将大数据中非结构化数据(例如视频和电子邮件)参与这一过程，与来自各种引擎 (追踪用户对你的品牌的评价)的新来源的信息，可以通过博客和用户论坛，然后将这些 信息与地理数据相关联，并结合现有结构化客户数据，从而获得强大的预测能力。

**2.** **缩短大数据分析时间**

大数据分析的一个很大优势是缩短预测时间，数据科学家曾经需要花几个月时间来建 立查询、模型以回答前瞻性业务问题，现在只需要几个小时就可以完成。这是因为利用大 数据技术可以实现信息在被优化或者关系化之前进行分析。再加上高级分析技术，使业务 经理在非常短的时间内询问和回答问题。之所以能这样做是因为大数据技术可以自动化建 模，并可以在无人值守的情况下执行。

**3.** **大数据预测分析和传统的数据仓库的不同**

大数据预测分析中包含了各种快速发展中的技术，简单用某一种技术将会比较困难。 传统的数据仓库系统通常从现有的关系型数据库中抓取数据。然而，超过80%的企业数据 是非结构化的，即无法用关系型数据库管理系统 (RDBMS) 来实现。 一般而言，非结构化 数据无法简单转化成关系型数据库中的数据。但是，企业现在希望从下述非结构化数据类 型中抽取有价值的信息。

① 邮件和其他形式的电子通信记录；

② 网站上的资料，包括点击量和社交媒体相关的内容；

③ 数字视频和音频；

④ 设备以及物联网产生的数据 (RFID 、GPS 、传感器产生的数据，日志文件等)。

在大数据分析中，查看多种大量的数据类型十分必要，这代表了各种重要的新信息源。

并且随着每年非结构化数据存储总量较结构化数据增长率高出10～50倍，从业务角度看， 这些数据也变得更为重要。但是，传统的数据仓库技术对非结构化数据的处理根本无法满 足大数据的需求。所以，存储管理人员更快地跟随技术发展，更新自己的技术和知识结构， 提高自己对大数据的管理和分析能力。

**11.6.9** **舆情监测与分析**

舆情监测分析并不是一个新鲜事物，古代的《邱报》算是今日舆情监测分析的鼻祖。





**大数据基础与应用**

**1.** **舆情监测内容**

监测内容主要包括对主流门户网站、国内外主流论坛、主流媒体、博客微博、主流搜 索引擎(如百度、谷歌等)等站点进行全景扫描，对相关刑事、民事、行政案件与信息进 行全面收集、精确分析、清晰归类和个性统计等工作，最终实现负面信息及时发现、重大 事件实时跟踪，自动生成各种统计报告，敏感信息及时预警等。监测内容的更详细说明如 下所述。

(1)网民披露监测

论坛、博客和微客现已成为网民表达自我想法、披露各种现象的最有效平台，个体都 已成为新闻的发布者、新闻的传播者和新闻的评论者。通过本系统可以第一时间发现网民 所发布的内容，如果是与法律有关问题，就会自动收录进系统。如果是重大负面事件，如 各种法律案件行为等，系统就会自动预警给单位相关负责人，可以第一时间掌握事件的实 际情况，通过实地考察迅速制订解决方案，第一时间给予妥善解决，最终获得广大人民群 众的认可和爱戴，完成人民所赋予的神圣使命。

(2)热点舆情监测

针对重大事件进行不间断实时监测，实现对热点事件设置监测关键词、监测周期、监 测范围、实时对重大事件相关信息进行汇总，自动生成传播趋势图、相关人员统计图、相 关机构统计图等各种图表。

(3)单位舆情监测

支持单位领导及相关部门为主题词的舆情信息采集，随时进行舆情信息监测。

**2.** **舆情分析概述**

舆情分析就是根据特定问题的需要，对针对这个问题的舆情进行深层次的思维加工和 分析研究，得到相关结论的过程。

(1)舆情分析的方法

① 内容分析法。内容分析法是一种对信息内容做出客观、系统的定量分析的专门方法， 其目的是弄清或测验信息中本质性的事实和趋势。提示信息所含有的隐性情报内容，对事 物发展做情报预测。

② 实证分析法。实证分析法是通过分析大量案例和相关数据后试图得出某些结论的一 种常见研究方法。

(2)舆情分析的注意点

对舆情的分析首先要明确事件或话题本身所处的阶段， 一般分为引发期、酝酿期、发 生期、发展期、高潮期、处理期、平息期和反馈期等不同阶段。其次，应该在分析某一舆 情热点之前对其进行科学的类型界定。热点事件一般主要分为突发自然灾害事件、生产安 全事故、群体性事件、公共卫生事件、公权力形象事件、司法事件、经济民生事件、社会 思潮事件、境外涉华突发事件等。

**3.** **舆情分析的价值**

舆情分析具有重要的价值。在某个事件发生后，广大群众会通过各种途径了解到事情





**第11章** **大数据分析**

的真相，随后而来的便是纷纷评论，或支持或反对，或理性或感性，或热情参与或冷眼旁 观。当一种论调得到大家的认同后，舆情甚至可以对事件的走向产生重大的影响。而一旦 有心人可以从这舆情中分析出些什么,得到些什么,那么就可以做出一个正确的决定。舆 情分析就是一个风向标。

**4.** **舆情报告图表制作**

由于图表与列表能够清晰、直观、简洁、深刻、形象地以可视化方式表现舆情事件， 因此其制作的科学规范化特别重要。

① 一般常见的分析方法有连续接近法、举例说明法、比较分析法和流程图法等。应该 根据舆情事件本身的特征科学选择图表，如趋势图、比例饼图、百分比柱图、流程图等， 表格的设计则要简洁实用，科学高效。

② 舆情分析图表在数据来源上要注意区分传统媒体、网络论坛、博客、问答网站等。

③ 在媒体类型上注意区分媒体历史形态、媒体控制类型与地域类型。

④ 分析舆情言论主体的身份特征。

⑤ 在对各种观点做具体而微的定性分析，为研究结果做结论时应该注意材料之间的异 同，避免为了结论的独立精确而牺牲材料的丰富性，应该兼顾赞扬、支持、中立、不关心、 反对、谴责等不同态度，以防止观点遗漏导致分析结论偏颇。立体化、多层次、客观地反 映我国转型期多元化社会的不同利益诉求，为决策和研究提供科学全面的参考。

**5.** **大数据舆情分析的目标与特点**

大数据舆情分析的目标是要从复杂的大数据舆情中找出规律，也就是说，从网上及时 发现舆情话题的热点，通过对舆情样本进行科学收集、抽样、统计、汇总、分析，准确把 握舆情现状，并在此基础之上，对其发展走势给出预测，提出应对与处理的意见和建议。 大数据舆情分析的特点如下。

(1)快捷

舆情分析就是从网上找线索，找信息，然后进行加工处理。所以要快捷，这是因为网 上信息传播速度快，分析师要能够跟上节奏。

(2)准确

除了快捷之外，还要准确。在网上的信息很复杂而庞大，必须眼光独到，找到的是最 有价值与代表性的样本，这是一项艰难的工作。

(3)专业

专业是指专业的分析、研究与判断，科学的应对和处置方法，要有理论基础和实践经验。

要追踪所服务部门的舆情态势，例如发现最近的舆情热点，通过科学的抽样分析总结 出趋势规律，进行文字描述，并辅以图表，提出意见，从而生成舆情简报，同时通过各种 方式进行预警和汇报。

随着社交媒体和移动终端的飞速发展，尤其是社会媒体和移动终端的飞速发展，信息的 传播速度越来越快，参与讨论的网民也越来越多，导致大数据的产生。仅依靠传统的个人上 网方式不能完成及时有效地发现舆情、了解舆情全貌，因此，大数据舆情职业化应运而生。





大数据基础与应用

**11.7** **大数据分析应用**

随着大数据的出现，大数据分析技术得到了迅速的应用，下面介绍较典型的高价值大 数据应用。

**11.7.1** **为客户提供服务**

理解客户、定位客户，以及为客户提供服务，这是现在最典型的大数据应用领域之一。 企业热衷于收集社交媒体数据、浏览器日志、文本分析和传感器数据，使用大数据能够更 好地了解客户以及他们的行为和喜好，进而更全面地了解他们的客户。在大多数情况下， 总的目标是创建预测模型。例如通过应用大数据，公司可以更好地预测客户流失、可以更 好地预测哪些产品将会热卖，汽车保险公司能够了解其客户的驾驶水平。电力公司利用 Hadoop 分析来自智能电表的数据，这些智能电表可以自动完成计费功能。该公司还可以通 过收集输电线路上任意的电流波动信息。如果收集到这些信息并且能够描绘出电流变化图， 那么就可以在某个地方的变压器可能出现故障之前找到它，或者当发生停电事故时，会引 起电流的波动，公司就可以探测到波动之处，在用户打电话求助之前就采取行动。电力公 司还可以利用大数据技术来改善为客户所提供的服务，并通过电网监控、问题检测和对电 网进行微调等降低运营成本，但是这可能需要对某些正在老化的基础设施进行重大升级。

**11.7.2** **优化业务流程**

大数据也越来越多地用于优化业务流程。通过利用从社交媒体数据、网络搜索趋势 以及天气预报挖掘出的预测信息，零售商能够优化其库存。其中广泛应用大数据分析的 业务流程是供应链或配送路线优化。在这方面，地理定位或无线电频率识别传感器被用 来追踪货物或送货车，并通过整合实时交通数据来优化路线。人力资源业务流程也能够 通过使用大数据分析来改进。这包括优化人才招聘，以及使用大数据工具衡量公司文化 和人员参与度。

**11.7.3** **改善生活**

大数据不仅适用于企业和政府，也适用于每个人。现在可以利用可穿戴设备(例如智 能手表或智能手链)生成数据，这使我们可以追踪自己的热量消耗、睡眠模式等。还可以 利用大数据分析来寻找爱情，大多数网上交友网站都使用大数据工具和算法来帮助我们寻 找最合适的对象。

大数据分析的计算能力使我们能够在几分钟内解码整个DNA, 并让可以找到新的治疗 方法，同时更好地理解和预测疾病模式。就像所有人能够受益于智能手表和可穿戴设备产





**第** **1** **1** **章** **大数据分析**

生的数据一样，大数据同样可以帮助病人更好地治病。未来的临床实验将不会仅限于小样 本，而是将服务于每个人。大数据技术已经被用来监护早产婴儿以及患病婴儿。通过记录 和分析每次心跳以及呼吸模式，医生现在可以在任何身体不适症状出现之前预测24小时的 情况。这样，医生就可以更早地救助患病婴儿。

大数据还被用来改善城市。例如，它让城市可以基于实时交通信息、社交媒体和天气 数据来优化交通情况。很多城市正在试点大数据分析技术，试图转变为智能城市，将交通 基础设施和公共设施程序都加入进来。

**11.7.4** **提高体育成绩**

现在很多体育运动都已经开始采用大数据分析技术。例如用于网球鼻塞的 IBM SlamTracker 工具，使用视频分析来追踪足球或棒球比赛中每个球员的表现。而运动器材中 的传感器技术使我们可以获得比赛的数据以及如何改进的方案。很多精英运动队还追踪比 赛环境外运动员的活动，通过使用智能技术来追踪其营养状况及睡眠，甚至通过社交对话 来监控其情感状况。

**11.7.5** **优化机器和设备性能**

大数据分析还可以让机器和设备变得更加智能和自主化。例如，大数据技术可用来运 行自驾车。配有相机、 GPS 以及强大的计算机和传感器，未来在道路上安全驾驶的汽车不 需要人类的干预。大数据技术还可以用来优化智能电网。使用大数据技术也可以优化计算 机和数据仓库的性能。

**11.7.6** **改善安全和执法**

大数据被广泛应用于提高安全和执法过程。使用大数据分析来对抗恐怖主义活动，甚 至用来监控我们的生活。其他企业则使用大数据技术来检测和阻止网络攻击。警察还可以 使用大数据技术来捉拿罪犯，预测犯罪活动，检测欺诈性交易等。

**11.7.7** **金融交易**

大数据在金融行业的应用主要是在金融交易领域。高频交易是大数据应用比较多的领 域。其中，大数据算法被用来做出交易决定。现在，大多数股权交易都是通过大数据算法 进行，这些算法越来越多地开始考虑社交媒体网络和新闻网站的信息，在几秒内做出买入 和卖出的决定。

**11.7.8** **电信业务**

电信企业可以通过对大数据的应用，寻找更多、更好的用户，发现降低运营成本的方





**大数据基础与应用**

法和途径。在电信业务的大数据分析中，要求能够支持快速查询、实时数据处理等一系列 需求，呈现了大数据的分析处理复杂性，主要内容如下。

● 高效导入和清洗大数据；

● 实行分布式处理大数据；

● 在线分析处理与批量计算大数据；

● 需要利用采样，在精度损失较少的情况下大幅降低处理数据量；

● 为提升处理速度，需要优化硬件。

**11.7.9** **销售**

销售可来自其系统自动的产品推荐。通过客户分类、测试统计、行为建模、投放优 化四步，运营客户的行为数据带来竞争优势。在零售业有非常多的需要分析数据的需求， 但是按照传统手段是无法想象的，比如社交媒体网站、DVR 设备和日用品商店的会员卡 数据等。这个行业的数据如此的庞大和复杂，利用传统的数据库手段根本不可能进行分 析，所以零售商们转向大数据平台。同样，大数据技术也给市场营销公司带来革命性的 变化。公司所拥有的会员客户数据库十分庞大，规模有2.5PB, 其中包括了1.9亿家日用 品商店多年来的历史销售数据。它的最大的一个数据库就有令人难以置信的4.25 亿行数 据，公司每天需要在这个数据库中管理大约6.25亿行数据。通过分析这些数据，可帮助 一些主要的消费品制造商和大型连锁超市预测：消费者可能会购买什么商品，谁会对新 商品感兴趣等。

**11.8** **大数据分析平台与工具**

大数据分析具有很强的实践性，应该掌握一门数据分析语言， R 语言是数据分析的常 用语言。R 软件是免费、开源、专业的统计分析软件，功能强大，是实践和领会大数据建 模的有力途径。为了强调平台和工具的作用，下面介绍了较常用的大数据平台和工具。

**11.8.1** **大数据分析平台**

大数据分析平台应该具备如下主要特性。

● 可以快速迭代开发，不断更新功能；

● 存储所有数据，包括结构性数据、非结构形式、半结构性数据；

● 可以进行各种分析；

● 不但专业人员可以使用，而且业务人员也可以使用。

一个关键的大数据分析平台还应该可以接入与兼容不同的开发工具，实现数据分析的 图形化，并且可以在时间、地理空间等维度进行查询，接收各种不同的数据源，能够使用 各种不同的工具进行分析。





第11章 大数据分析

在大数据处理平台上，可以提供数据清洗、数据分析、数据挖掘、数据可视化的一系 列大数据分析功能。

**1.Hadoop 分布式计算平台**

Hadoop 是一个能够让用户轻松架设和使用的分布式计算平台。用户可以轻松地在 Hadoop 上开发和运行处理海量数据的应用程序，主要有高可靠性、高扩展性、高效性、高 容错性的优点。Hadoop 带有用Java 语言编写的框架，因此运行在 Linux 平台上。Hadoop 上的应用程序也可以使用 C++ 等其他语言编写。

**2.Storm**

Storm 是自由的开源软件， 一个分布式的、容错的实时计算系统。Storm 可以非常可靠 地处理庞大的数据流，用于处理 Hadoop 的批量数据。Storm 支持许多种编程语言。

Storm 有许多应用领域：实时分析、在线机器学习、不停顿的计算、分布式远程调用 协议。

**3.Apache Drill**

为了帮助企业用户寻找更为有效、加快 Hadoop 数据查询的方法，Apache 软件基金会 发起了一项名为 Drill的开源项目。Apache Drill 实现了Google's Dremel。该项目将会创建 出开源版本的谷歌 Dremel Hadoop 工具，而 Drill 将有助于 Hadoop 用户实现更快查询海量 数据集的目的。Drill 项目实现海量数据集的分析处理，包括分析抓取 Web文档、跟踪安装 在 Android Market上的应用程序数据、分析垃圾邮件、分析谷歌分布式构建系统上的测试 结果等。通过开发 Apache Drill 开源项目，组织机构将有望建立 Drill 所属的 API 接口和灵 活强大的体系架构，从而帮助支持广泛的数据源、数据格式和查询语言。

**11.8.2** **大数据分析的工具**

下面介绍几种大数据分析的工具。

**1.EMC Greenplum 统一分析平台**

EMC Greenplum 统一分析平台 (UAP) 是一个单一软件平台，数据团队和分析团队可 以在该平台上无缝地共享信息、协作分析。EMC Greenplum 统一分析平台包括 ECM Greenplum关系数据库、EMC Greenplum HD Hadoop 发行版和 EMC Greenplum Chorus。

EMC 为大数据开发的硬件是模块化的 EMC 数据计算设备 (DCA), 它能够在一个设 备里面运行并扩展Greenplum 关系数据库和Greenplum HD节点。DCA 提供了一个共享的 指挥中心界面，管理员可以监控、管理和配置 Greenplum 数据库和 Hadoop 系统性能及容 量。随着Hadoop 平台日趋成熟，分析功能将迅速增强。

**2.Biglnsights 和** **BigCloud**

IBM 推出了InfoSphere BigI 云版本的InfoSphere BigInsights, 使组织内的任何用户都 可以做大数据分析。云上的 BigInsights 软件可以分析数据库里的结构化数据和非结构化数



**大数据基础与应用**

据，使决策者能够迅速将洞察转化为行动。

**3.Informatica 9.1**

HParser 支持灵活高效地处理 Hadoop 中的任何文件格式，为 Hadoop 开发人员提供了 即开即用的解析功能，以便处理复杂而多样的数据源，包括日志、文档、二进制数据或层 次式数据，以及众多行业标准格式(如银行业的 NACHA 、支付业的 SWIFT 、金融数据业 的 FIX 和保险业的ACORD) 。正如数据库内处理技术加快了各种分析方法，Informatica 同 样将解析代码添加到 Hadoop 中，以便充分利用所有这些处理功能。

**4.Vertica 数据分析平台**

惠普发布的 Vertica 5.0 是能提供高效数据存储和快速查询的列存储数据库实时分析平 台。该数据库还支持大规模并行处理。惠普随即推出了基于x86 硬件的 HP Vertica。通过 MPP 的扩展性可以让Vertica 为高端数字营销、电子商务客户分析处理的数据达到PB 级。惠普展 示了一款 Vertica Analytics Appliance 设备，它是惠普融合基础架构中的一款全集成技术栈。

**5.Oracle Big Data Appliance**

甲骨文的 Big Data Appliance 集成系统包括 Cloudera 的 Hadoop 系统管理软件和支持服 务Apache Hadoop 和 Cloudera Manager。甲骨文把 Big Data Appliance 看作包括 Exadata 、 Exalogic 和 Exalytics In-Memory Machine 的建造系统。Oracle 大数据机是一个软、硬件集 成系统，该大数据机采用Oracle Linux操作系统，并配备 Oracle NoSQL 数据库社区版本和 Oracle HotSpot Java 虚拟机。Big Data Appliance 为全架构产品，每个架构864GB 存储，216 个 CPU 内核，648TB RAW 存储，每秒40GB 的 InifiniBand 连接。

**6.** **微软** **PDW**

2011年初，微软发布了SQL Server R2 Parallel Data Warehouse(PDW, 并行数据仓库)。 PDW 使用了大规模并行处理来支持高扩展性，并可以帮助客户扩展部署数百TB 级别数据 的分析解决方案。微软目前已经开始提供 Hadoop Connector for SQL Server Parallel Data Warehouse 和 Hadoop Connector for SQL Server社区技术预览版本的连接器。该连接器是双 向的，可以在 Hadoop 和微软数据库服务器之间向前或者向后迁移数据。微软现在已推出 基于Azure 云平台的测试版 Hadoop 服务，将推出与 Windows 兼容的基于 Hadoop 的大数据 解决方案 (Big Data Solution), 这是微软 SQL Server 2012 版本的一部分。

**7.** **亚马逊弹性MapReduce**

亚马逊在2009年就推出了亚马逊弹性 MapReduce(Amazon Elastic MapReduce),亚马 逊对 Hadoop 的需求和应用了如指掌。弹性 MapReduce 是一项能够迅速扩展的 Web 服务， 运行在亚马逊弹性计算云 (AmazonEC2) 和亚马逊简单存储服务 (Amazon S3) 上。这种 云面对数据密集型任务，比如互联网索引、数据挖掘、日志文件分析、机器学习、金融分 析、科学模拟和生物信息学研究，用户需要多大容量，立即就能配置到多大容量。除了数 据处理外，用户还可以使用 Karmasphere Analyst的基于服务的版本，Karmasphere Analyst





第11章大数据分析

是一种可视化工作区，用于在亚马逊弹性 MapReduce 上分析数据。用户还可以提取结果文 件，以便在数据库或者微软 Excel 或 Tableau 等工具中使用。

**8.Aster Data**

Teradata 是企业级数据仓库 (EDW) 的领导者，在数据库分析领域不断推陈出新。为 了发展结构化数据、半结构化数据和大部分非结构化数据领域，该公司要收购 Aster Data。 它是一家提供 SQL-MapReduce 框架的公司。Teradata 宣布了一项 Aster Data MapReduce 产 品的计划，它建立在以往产品同样的硬件平台之上，而且在 Teradata和 Aster Data 之间新 增了两种集成方法。Aster Data 是高级分析和管理各种非结构化数据领域的市场领导者和开 拓者。Aster Data为 Teradata 带来了大数据分析市场商，加之收购Aprimo 所获得的整合营 销管理 (Integrated Marketing Management) 能力，以及不断加大的核心数据仓库业务投资 力度，将为 Teradata的未来发展注入强劲动能。

**9.Google Dremel 数据分析系统**

Dremel 是 Google 的交互式数据分析系统。可以构建成规模上千的集群，处理 PB 级别 的数据。MapReduce 处理一个数据，需要分钟级的时间， Dremel 将处理时间缩短到秒级的 时间，可以作为MapReduce 的有力补充。Dremel 作为Google BigQuery 的 report 引擎，获 得了很大的成功。随着 Hadoop 的流行，大规模的数据分析系统已经越来越普及。数据分 析师需要一个能将数据运转的交互式系统，就可以非常方便快捷地浏览数据，建立分析模 型。Dremel 系统的主要特点如下。

(1)Dremel 是一个大规模系统

对于一个 PB 级别的数据集，将任务缩短到秒级，无疑需要大量的并发。磁盘的顺序 读速度约100MB/s, 那么在1s 内处理1TB 数据，表明至少需要有1万个磁盘的并发读。利 用廉价机器可以完成强大功能，但是机器越多，出问题概率越大，因此集群规模大，需要 有足够的容错考虑，保证整个分析的速度不被集群中的个别慢节点影响。

(2)Dremel 是 MapReduce 模型的交互式查询功能的补充

与 MapReduce一样，Dremel 也需要和数据运行在一起，将计算移动到数据上面。所以 它需要GFS 文件系统作为存储层。在设计之初， Dremel 并不是MapReduce 的替代品，它 只是可以执行分析非常快，在使用时，常常用它来处理 MapReduce 的结果集或者用来建立 分析原型。

(3)Dremel 的数据模型是嵌套的

互联网数据常常是非关系型的。Dremel 还需要有一个灵活的数据模型，这个数据模型 至关重要。Dremel 支持一个嵌套的数据模型，类似于Json 。而传统的关系模型，由于不可 避免地有大量的 Join 操作，在处理如此大规模的数据的时候，往往有心无力。

(4)Dremel 中的数据是用列式存储的

使用列式存储，分析的时候，可以只扫描需要的那部分数据，减少 CPU 和磁盘的访问 量。同时列式存储是可压缩的，使用压缩，可以综合CPU 和磁盘，发挥最大的效能。对于 关系型数据，如果使用列式存储，都很有经验。但是对于嵌套的结构，Dremel 也可以用列





大数据基础与应用

存储，非常值得我们学习。

(5)Dremel 结合了 Web 搜索和并行 DBMS 的技术

首先，利用了 Web 搜索中的查询树的概念，将一个相对巨大复杂的查询分割成较小较 简单的查询。其次，与并行 DBMS 类似，Dremel 可以提供一个 SQL-like 的接口，就像Hive 和 Pig 那样。

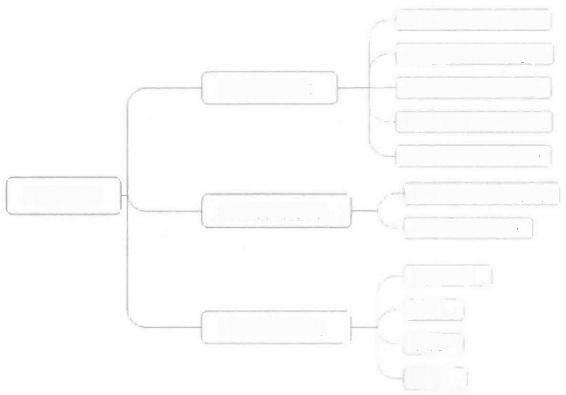
小结

随着软件工具、数据库技术、各种硬件设备的飞快发展，使得我们分析大数据成为可 能。大数据分析从大数据中提取、挖掘对业务发展有价值的潜在的知识，找出趋势，为决 策层提供有力依据，为产品或服务发展方向起到积极作用，有力推动企业内部的科学化、 信息化管理。



**第12章** **大数据挖掘**

本章主要内容



数据挖掘的几个概念

数据挖掘对象与过程

数据挖掘的常用方法

数据挖掘的几个问题

数据挖掘的经典算法

大数据挖掘关键技术

大数据挖掘技术

大数据挖掘策略

市场营销

销售

大数据挖掘应用

物流

CRM

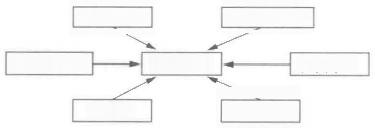
数据挖掘概述

大数据挖掘

数据挖掘是指在大量的数据中挖掘出信息，通过认真分析来揭示数据之间有意义的联 系、趋势和模式。而数据挖掘技术就是指为了完成数据挖掘任务所需要的全部技术。金融、 零售等企业已广泛采用数据挖掘技术，分析用户的可信度和购物偏好等。大数据研究采用 数据挖掘技术，但是数据挖掘中的短期行为较多，多数是为某个具体问题研究应用技术， 还无统一的理论。传统的数据挖掘技术在数据维度和规模增大时，所需资源呈现指数级增 长，所以对PB 级以上的大数据还需研究新的方法。

**12.1** **数据挖掘概述**

数据挖掘是近年来伴随数据库系统的大量建立和万维网的广泛应用而发展起来的一门 技术。数据挖掘是交叉性学科，它是数据库技术、机器学习、统计学、人工智能、可视化 分析、模式识别等多门学科的融合，如图12-1 所示。



模式识别 可视化分析

数据库技术 数据挖掘 机器学习

统计学 肝算机网路

图12-1 数据挖掘是多学科的融合



**大数据基础与应用**

**12.1.1** **数据挖掘的几个概念**

**1.** **数据挖掘**

数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐 含其内的、人们事先所不知的，但又是有潜在价值的信息和知识的过程。几点说明如下。

① 数据挖掘涉及数据融合、数据分析和决策支持等内容。

② 数据源必须是真实的、大量的、含有噪声的、用户感兴趣的数据。

③ 发现的知识要可接受、可理解、可运用，并不要求发现放之四海而皆准的知识，仅 支持特定的问题。

④ 数据是知识的源泉，将概念、规则、模式、规律和约束等视为知识，这就好像从矿 石中采矿或淘金一样，从数据中获取知识。

⑤ 原始数据可以是结构化数据，如关系型数据库中的数据等，也可以是非结构化数据， 如文本、图形和图像等，还可以是半结构化数据，如网页等。

⑥ 挖掘知识的方法可以是数学的方法，也可以是非数学的方法；可以是演绎的方法， 也可以是归纳的方法。

⑦ 挖掘的知识具有应用的价值，可以用于信息管理、查询优化、决策支持和过程控制 等，还可以用于数据自身的维护。

⑧ 数据挖掘是一门交叉学科，将人们对数据的应用从低层次的简单查询，提升到从数 据中挖掘知识，提供决策支持。在需求推动下，不同领域的研究者，尤其是数据库技术、人 工智能技术、数理统计、可视化技术、并行计算等方面的知识融合后，形成新的研究热点。

数据的挖掘首先是搜集数据，数据越丰富越好，数据量越大越好，只有获得足够的高 质量的数据，才能获得确定的判断，才能产生认知模型，这是量变到质变的过程。由此产 生经验，经验的积累就能产生有价值的判断。认知模型是渐进发展的模型，当认识深入以 后，将产生更加抽象的模型与许多猜想，通过猜想再扩展模型，从而达到深度学习和深度 挖掘。

**2.** **数据挖掘分类**

数据挖掘可以分为两类：直接数据挖掘和间接数据挖掘。

(1)直接数据挖掘

直接数据挖掘的目标是利用可用的数据建立一个模型，利用这个模型对剩余的数据，

对一个特定的变量(可以理解成数据库中表的属性，即列)进行描述。分类、估值、预测 属于直接数据挖掘。

(2)间接数据挖掘

间接数据挖掘目标中没有选出某一具体的变量，用模型进行描述，而是在所有的变量 中建立起某种关系。相关性分组或关联规则、聚类、描述和可视化以及复杂数据类型挖掘 属于间接数据挖掘。





**第12章** **大数据挖掘**

**3.** **数据挖掘技术**

数据挖掘技术是数据挖掘方法的集合，数据挖掘方法众多。根据挖掘任务可将数据挖 掘技术分为预测模型发现、聚类分析、分类与回归、关联分析、序列模式发现、依赖关系 或依赖模型发现、异常和趋势发现、离群点检测等。根据挖掘对象可分为关系数据库、面 向对象数据库、空间数据库、时态数据库、文本数据源、多媒体数据库、异质数据库、遗 产数据库以及环球网 Web。根据挖掘方法可分为机器学习方法、统计方法、神经网络方法 和数据库方法。机器学习方法中，可细分为归纳学习方法(决策树、规则归纳等)、基于范 例学习、遗传算法等。统计方法中，可细分为回归分析(多元回归、自回归等)、判别分析 (贝叶斯判别、费歇尔判别和非参数判别等)、聚类分析(系统聚类、动态聚类等)、探索 性分析(主元分析法、相关分析法等)等。神经网络方法中，可细分为前向神经网络 (BP 算法等)、自组织神经网络(自组织特征映射、竞争学习等)等。数据库方法主要是多维数 据分析或 OLAP 方法，另外还有面向属性的归纳方法。

数据挖掘应用了来自其他一些领域的思想与算法，主要包括：

① 统计学的抽样、估计和假设检验。

② 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。

③ 最优化、进化计算、信息论、信号处理、可视化和信息检索。

其他一些领域的技术也起到重要的支撑作用，需要数据库系统提供有效的存储、索引 和查询处理支持。高性能计算技术、并行计算技术、分布式技术也能帮助处理数据，当数 据不能集中到一起处理时更是至关重要。

**12.1.2** **数据挖掘对象与过程**

**1.** **数据挖掘对象**

数据挖掘对象是指用于挖掘知识的数据来源，主要的挖掘对象有关系数据库、面向对 象数据库、数据仓库、文本数据源、多媒体数据库、空间数据库、时态数据库、异质数据 库、NoSQL 数据库以及Internet 等。这里主要介绍下述几种。

(1)文本挖掘

文本挖掘是将不同的文档进行比较之后，进行文档重要性和相关性排列，整理出文档 的模式和趋势。文本挖掘的处理过程包括对大量文档集的内容进行预处理、特征提取、结 构分析、文本摘要、文本分类、文本聚类、文本关联等。其中不仅需要进行结构化和非结 构化文档数据的处理，而且还需要处理复杂的语义关系。对于非结构化数据的挖掘存在两 种解决方法， 一种是开发新的非结构化数据挖掘算法，另一种是将非结构化数据转变成结 构化数据，利用已有的数据挖掘算法进行挖掘，这是目前广泛应用的方法。

(2)Web 挖掘

Web挖掘是指从与WWW 有关联的资源及行为中获取有用的模式和隐含的信息。Web 中含有大量信息和超链接信息、Web 页面的访问和使用信息，是进行 Web 挖掘的重要资源。 Web 挖掘包括以下三方面。





**大数据基础与应用**

① 在文档内容或文档描述中获取知识的过程，称之为 Web内容挖掘。

② 从 WWW的组织结构及链接关系中推导知识，称之为 Web 结构挖掘。 ③ 从 Web的访问记录中抽取感兴趣的模式，即 Web 使用记录的挖掘。

(3)多媒体数据的挖掘

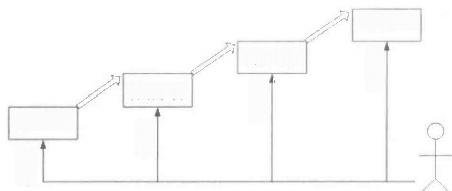
多媒体数据类型包括图像、视频、音频时空数据和超文本等，隐藏了大量的有价值的 知识。多媒体数据的挖掘是综合分析大量多媒体数据的视听特性与语义，利用多媒体的时 间、空间、视觉特性、视听对象以及运动特性，挖掘出具有一定价值的、能够理解的知识 模式，找出实践的趋势以及关联性。

(4)时空数据挖掘

时空数据挖掘是指从海量、高维、高噪声和非线性的时空数据中提取隐含的、人们事 先不知的、潜在的有用信息及知识的过程。时间维度和空间维度的存在增加了时空数据挖 掘的复杂性。通常将时空数据挖掘分为时空模式发现、时空聚类、时空异常检测、时空预 测和分类等。

**2.** **数据挖掘过程**

数据挖掘是通过分析每个数据，从大量数据中寻找其规律的技术，虽然数据挖掘的过 程随着不同领域的应用而有所区别，但一般认为有问题定义、数据准备、规律寻找和规律 表示4个步骤，如图12-2所示。数据准备是从相关的数据源中选取所需的数据并整合成用 于数据挖掘的数据集；规律寻找是用某种方法将数据集所含的规律找出来；规律表示是用 用户可理解的方式，如可视化方式，表示已找出的规律。



规律表示

规律寻找

数据准备

问题定义

指 导

指 导

指 导

指导

图12-2 数据挖掘过程

数据挖掘更详细的步骤如下。

(1)定义问题，确定业务对象

清晰地定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结 构是不可预测的，但要探索的问题应是有预见性的。

(2)数据准备

数据准备包括数据的选择、数据的预处理和数据的转换。

● 数据的选择：在大型数据库和数据仓库目标中提取数据挖掘的目标数据集，搜索 所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用 的数据。

● 数据的预处理：研究数据的质量，为进一步的分析做准备，并确定将要进行的挖



**第12章** **大数据挖掘**

掘操作的类型。进行数据再加工，包括检查数据的完整性及数据的一致性、去噪 声，填补丢失的域，删除无效数据等。

● 数据的转换：将数据转换成一个分析模型。该分析模型是针对挖掘算法建立的。 建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

(3)数据挖掘

根据数据功能的类型和数据的特点选择相应的算法，在净化和转换过的数据集上进行 数据挖掘。

(4)结果分析

对数据挖掘的结果进行解释和评价，转换成能够最终被用户理解的知识。对所得到的 经过转换的数据进行挖掘，除了选择合适的挖掘算法外，其余一切工作都能自动地完成。

(5)知识的运用

将分析所得到的知识集成到业务信息系统的组织结构中，实现知识的应用。

**3.** **数据挖掘工作量**

数据挖掘是一个完整的过程，该过程从大型数据库中挖掘先前未知的、有效的、可实 用的信息，并利用这些信息做出决策或丰富知识。

在数据挖掘中，被研究的业务对象驱动了整个数据挖掘过程，也是检验最后结果和指 引分析人员完成数据挖掘的根据。当然，整个过程中还会存在步骤间的反馈。数据挖掘的 过程并不是自动的，绝大多数的工作需要人工完成。图12-3给出了各步骤在整个过程中的 工作量之比。从图中可以看出，60%的时间用在数据准备上，这说明数据挖掘对数据的严 格要求，而后挖掘工作仅占总工作量的10%。

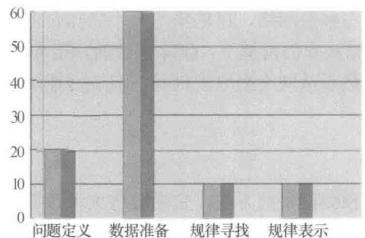


图12-3 数据挖掘过程工作量比例

数据挖掘过程的分步实现需要不同专长的人员，大体可以分为以下三类。

(1)业务分析人员

要求精通业务，能够解释业务对象，并根据各业务对象确定出用于数据定义和挖掘算 法的业务需求。

(2)数据分析人员

精通数据分析技术，并对统计学有较熟练的掌握，有能力把业务需求转化为数据挖掘 的各步操作，并为每步操作选择合适的技术。



**大数据基础与应用**

(3)数据管理人员

精通数据管理技术，并从数据库或数据仓库中收集数据。

可见，数据挖掘涉及较多领域的知识，需要多领域专家合作完成，数据挖掘过程是需 要反复进行的过程，不断地逼近事物的本质，不断地优化问题的解决方案。

**12.1.3** **数据挖掘的常用方法**

下面介绍几种最常用的数据挖掘方法。

**1.** **神经网络方法**

神经网络是基于生理学建立的智能仿生系统模型，具有良好的鲁棒性、自组织性、自 适应性、并行处理、分布存储和高度容错性等特性，非常适合解决数据挖掘的问题，因此， 近年来越来越受到重视。典型的神经网络模型有感知机、BP 反向传播模型、函数型网络等。 前馈式神经网络模型主要用于分类、预测和模式识别； Hopfield 的离散模型和连续模型分 别用于联想记忆和优化计算的反馈式神经网络模型； ART 模型和Koholon 模型主要用于聚 类的自组织映射。

**2.** **遗传算法**

遗传算法是一种基于生物自然选择与遗传机理的随机搜索算法，是一种仿生全局优化 方法。遗传算法具有的隐含并行性、易于和其他模型结合等性质使得它在数据挖掘中被加 以应用。现已成功地开发了一个基于遗传算法的数据挖掘工具，利用该工具对两架飞机失 事的真实数据库进行了数据挖掘实验，结果表明遗传算法是进行数据挖掘的有效方法之一。 遗传算法的应用还体现在与神经网络、粗糙集等技术的结合上。例如，利用遗传算法优化 神经网络结构，在不增加错误率的前提下，删除多余的连接和隐层单元；用遗传算法和 BP 算法结合训练神经网络，然后从网络提取规则等。但遗传算法较复杂，较早陷于局部极小 的收敛问题还没有解决。

**3.** **决策树**

决策树是一种常用于预测模型的方法，它通过将大量数据有目的地分类，从中找到一 些有价值的、潜在的信息。主要优点是描述简单，分类速度快，特别适合于大规模的数据 处理。最有影响和最早的决策树方法是著名的基于信息熵的 ID3 算法。它的主要问题是： ID3 是非递增学习算法； ID3 决策树是单变量决策树，复杂概念的表达困难；同性间的相互 关系强调不够；抗噪性差。针对上述问题，出现了一系列的改进算法。

**4.** **粗糙集方法**

粗糙集理论是研究不精确、不确定知识的数学工具。粗糙集方法的优点是不需要给出 额外信息、简化输入信息的表达空间、算法简单而易于操作。粗糙集处理的对象是类似于 二维关系表的信息表。关系数据库管理系统和数据仓库管理系统为粗糙集的数据挖掘奠定 了坚实的基础。但粗糙集的数学基础是集合论，难以直接处理连续的属性，而现实信息表





**第12章** **大数据挖掘**

中连续属性是普遍存在的，所以连续属性的离散化制约了粗糙集理论的实用化。

**5.** **覆盖正例排斥反例方法**

覆盖正例排斥反例方法用覆盖所有正例、排斥所有反例来寻找规则。首先在正例集合 中任选一个种子，到反例集合中逐个比较。与字段取值构成的选择子相容则舍去，相反则 保留。按此思想循环所有正例种子，将得到正例的规则，即选择子的合取式。

**6.** **统计分析方法**

在数据库字段项之间存在两种关系：函数关系(能用函数公式表示的确定性关系)和 相关关系(不能用函数公式表示，但仍是相关确定性关系),对它们的分析可采用统计学方 法，即利用统计学原理对数据库中的信息进行分析，主要内容介绍如下。

① 统计分析：求解数据集中的最大值、最小值、总和、平均值等。

② 回归分析：用回归方程来表示变量间的数量关系。

③ 相关分析：用相关系数来度量变量间的相关程度。

④ 差异分析：从样本统计量的值得出差异来确定总体参数之间是否存在差异等。

**7.** **模糊集方法**

模糊集方法是利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别 和模糊聚类分析。系统的复杂性越高，模糊性越强， 一般模糊集合理论是用隶属度来描述 模糊事物的特性。

**8.** **离群点检测**

离群点又称异常点、孤立点，离群点检测的任务是识别特征显著不同于其他数据的观 测值，离群点检测算法的目标是发现真正的离群点，同时避免将正确的对象标注为离群点。

离群点检测的定义如下：给定一个有N 个数据点或对象的数据集和期望的离群点数K, 找出与数据集中其余数据显著不同的、异常的或不一致的前K 个对象。挖掘离群点问题可 以看作下述两个子问题， 一个是定义在给定的数据集中，什么样的数据可以认为是不一致 的；另一个是找到一个有效的方法来挖掘所定义的离群点。

(1)离群点产生的原因

存在多种原因导致离群点的产生，归纳如下。

① 欺诈、入侵、不寻常的实验结果数据。

② 自然发生，表示了数据集的数据分布特征，如气候变化、基因突变等。

③ 数据测量和收集出现了误差，主要有人为错误、测量设备故障或存在噪声等。 (2)离群点分类

离群点分类如图12-4所示。

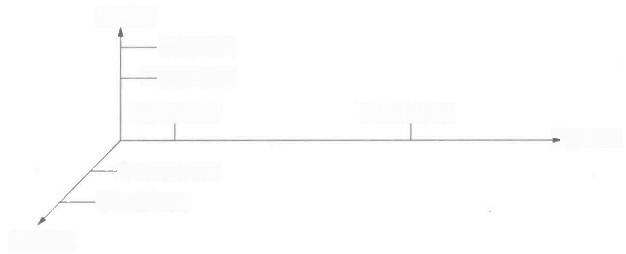
(3)离群点检测

常用的离群点检测方法如下。





大数据基础与应用



数据范围

全局离群点

—局部离群点

单变量离群点

→变量数目

一分类型离群点

-数值型离群点

数据类型

多变量离群点

图12-4 离群点分类

① 基于统计的离群点检测。在统计方法中，首先对已知的给定数据样本集假设一个概 率模型，例如泊松分布或正态分布，然后结合假设的模型进行不一致性检验，确定离群点。

② 基于距离的离群点检测。对象 O 是一个基于距离的孤立点，可以表示为 DB(p,d ), 其中p 和 d 是对象O 的参数，表示数据集S 中至少存在 p 部分与对象O 的距离大于 d, 也 就是说，将那些与给定对象的距离较大的对象作为距离孤立点。基于距离的离群点检测的 最大优点是减少了计算量。

③ 基于偏差的离群点检测。基于偏差的离群点检测是一种通过检测对象的特征来找出 不一致性数据的方法。如果一个对象偏离了给定对象的特征描述，那么此对象是不一致的， 这也表明偏差的含义是异常。常用的基于偏差的离群点检测方法是序列异常技术和 OLAP 数据立方体技术。序列异常方法模仿了人类可以从一系列类似的对象中识别出异常对象的 行为。OLAP 数据立方体方法是使用了数据立方体来辨识高维数据中的异常区域，这种方 法预先计算的是数据异常的度量，用于在数据集计算的所有层次上指导用户进行数据分析。 如果数据立方体中的一个单元值与基于统计模型的期望值显著不同，那么这个单元值就可 以认为异常。

**12.1.4** **数据挖掘的几个问题**

**1.** **数据挖掘与数据分析的区别**

数据挖掘与传统的数据分析，如查询、报表、联机应用分析等的本质区别是数据挖掘 是在没有明确假设的前提下去挖掘信息、发现知识。数据挖掘所得到的信息应具有先前未 知性、有效性和可实用性三个最基本的特征。

先前未知的信息是指该信息是预先未曾预料到的，即数据挖掘是要发现那些不能靠直 觉发现的信息或知识，甚至是违背直觉的信息或知识，挖掘出的信息越是出乎意料，就可 能越有价值。在商业应用中最典型的例子就是一家连锁店通过数据挖掘发现了婴儿使用的 尿不湿和啤酒之间的惊人联系。

**2.** **数据挖掘和数据仓库**

要将庞大的数据转换成有用的信息，必须先有效率地收集信息。简单地说，数据仓库





数据源  数据挖掘库

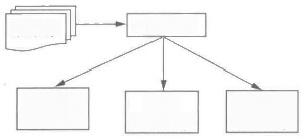


**第12章** **大数据挖掘**

就是搜集来自其他系统的有用数据，存放在一个整合的存储区内。所以数据仓库就是一个 经过处理整合，且容量特别大的关系型数据库，用以存储决策支持系统所需的数据，供决 策支持或数据分析使用。

数据仓库本身是一个非常大的数据库，它存储着由组织作业数据库中整合而来的数据， 特别是指事务处理系统 OLTP(On-Line Transactional Processing) 所得来的数据。将这些整 合过的数据置放于数据仓库中，而决策者则利用这些数据做决策，但是，这个转换及整合 数据的过程是建立一个数据仓库。因为将作业中的数据转换成有用的策略性信息是整个数 据仓库的重点。综上所述，数据仓库应该具有：整合性数据、详细和汇总性的数据、历史 数据、解释数据的数据。从数据仓库挖掘出对决策有用的信息与知识，是建立数据仓库与 使用数据挖掘的最大目的。换句话说，数据仓库应先行建立完成，数据挖掘才能有效率地 进行，因为数据仓库本身所含数据是干净(不会有错误的数据掺杂其中)、完备且经过整合 的。因此两者关系或许可解读为数据挖掘是从巨大数据仓库中找出有用信息的一种过程与 技术。

通常，要先把数据从数据仓库中传送到数据挖掘库或数据集市中进行数据挖掘，如 图12-5与图12-6 所示。但是，从数据仓库中直接进行数据挖掘的优点是数据仓库的数据 清理和数据挖掘的数据清理差不多，如果数据在导入数据仓库时已经清理，在做数据挖掘 时就没必要再清理一次，而且所有的数据不一致的问题都已经解决。



数据源

各分公司 数据集市

数据仓库

分析 数据集市

数据挖掘 数据集市

图12-5 数据挖掘库从数据仓库中导出 图12-6 数据挖掘库从事务数据库中得出

数据挖掘库可以是数据仓库的一个逻辑子集，而不必是物理上独立的数据库。如果数 据仓库的计算资源已经很紧张，那最好还是建立一个单独的数据挖掘库。当然，为了数据 挖掘也不必非得建立一个数据仓库，也就是说，数据仓库不是必需的。建立一个巨大的数 据仓库，把各个不同源的数据统一在一起，解决所有的数据冲突问题，然后把所有的数据 导入一个数据仓库内，需要较大的投入和较长的时间才能完成。如果只是为了数据挖掘， 也可以把一个或几个事务数据库导入到一个只读的数据库中，然后进行数据挖掘。

**3.** **数据挖掘和** **OLAP 的比较**

OLAP(Online Analytical Process) 是指由数据库所链接出来的在线分析处理。OLAP 是决策支持领域的一部分。数据挖掘与 OLAP 截然不同，数据挖掘用在产生假设， OLAP 则用于查证假设。简单来说，OLAP 是由使用者所主导，使用者先有一些假设，然后利用 OLAP 来查证假设是否成立，而数据挖掘则是用来帮助使用者产生假设。所以在使用OLAP 或其他 Query 的工具时，使用者是自己在做探索，但数据挖掘是用工具系统在帮助做探索。

如果分析工程师想找到导致贷款拖欠的原因，可能先做一个初始的假定，认为低收入





**大数据基础与应用**

的人信用度也低，然后用OLAP 来验证这个假设。如果这个假设没有被证实，可以去查看 那些高负债的账户，如果还不行，他也许要把收入和负债一起考虑， 一直进行下去，直到 找到需要的结果或放弃。也就是说， OLAP 分析师是建立一系列的假设，然后通过 OLAP 来证实或推翻这些假设来最终得到结论。OLAP 分析过程在本质上是一个演绎推理的过程。 但是如果分析的变量达到几十或上百个，那么再用 OLAP 分析验证这些假设将是一件非常 困难的事情。数据挖掘与 OLAP 不同的地方是，数据挖掘不是用于验证某个假定的模式(模 型)的正确性，而是在数据库中自己寻找模型。在本质上这是一个归纳的过程。例如， 一 个用数据挖掘工具的分析师需要找到引起贷款拖欠的风险因素。数据挖掘工具可能帮助找 到高负债和低收入是引起这个问题的因素，甚至还可能发现一些分析师从来没有想过或试 过的其他因素，如年龄等。数据挖掘与OLAP 具有互补性。在利用数据挖掘出来的结论采 取行动之前，也需要验证一下如果采取这样的行动会带来什么样的影响，那么 OLAP 工具 能回答这些问题。而且在知识发现的早期阶段， OLAP 工具还有其他用途，如探索数据， 找到对一个问题比较重要的变量，发现异常数据和互相影响的变量。这能更好地理解数据， 加快知识发现的过程。

数据挖掘常能够挖掘出超越归纳范围的关系，但 OLAP 仅能利用人工查询及可视化的 报表来确认某些关系。利用数据挖掘可以自动找出数据模型与关系的特性， OLAP 可以和 数据挖掘互补。

**4.** **数据挖掘和人工智能**

人工智能和统计分析的目标都是模式发现和预测，但数据挖掘不是为了替代传统的统 计分析技术，而是统计分析方法学的延伸和扩展。大多数的统计分析技术都基于完善的数 学理论和高超的技巧，预测的准确度令人满意，但却对使用者的要求很高。随着计算机系 统计算能力的不断增强，可以利用强大的计算能力，在足够多的数据支持下，应用计算智 能，几乎不用人的参与即可自动完成许多有价值的功能。

数据挖掘利用了统计和人工智能技术的应用程序，封装这些高深复杂的技术，不用人 的参与掌握这些技术也能完成同样的功能，这样可使人们将更多的精力专注于需要解决的 问题。

**5.** **数据挖掘和统计分析**

统计学通过机器学习来影响数据挖掘，而机器学习和数据库则是数据挖掘的两大支撑 技术。从数据分析的角度来看，绝大多数数据挖掘技术都来自机器学习领域，但机器学习 研究并不把海量数据作为处理对象，因此，数据挖掘要对算法进行改造，使得算法性能和 空间占用达到实用的地步。相对传统统计分析而言，数据挖掘有下列几项特性。

① 数据挖掘处理大量实际数据，无须过多专业的统计背景去使用数据挖掘的工具。

② 数据分析趋势是从大型数据库抓取所需数据并使用专用分析软件，数据挖掘的工具 更符合企业需求。

③ 从理论的基础点来看，数据挖掘和统计分析在应用上存在差别，数据挖掘的目的是 方便企业终端用户使用，而非用于统计学检测。



第12章 大数据挖掘

**6.Web 挖掘和数据挖掘的区别**

Web 挖掘是指网络数据挖掘。从堆积如山的由网络所获得的数据中找出让网站运作更 有效率的操作因素。Web 挖掘不只限于一般的日志文件分析，除了计算网页浏览率以及访 客人次外，还有网络上的零售、财务服务、通信服务、政府机关、医疗咨询、远距教学等， 只要由网络联结出的数据库大而完整，即可进行离线的分析与 Web 挖掘，或可整合离线及 在线的数据库，实施更大规模的模型预测与推算。Web 挖掘具有以下特性。

① 数据收集容易。

② 以交互式个人化服务。

③ 可整合外部来源数据。

利用数据挖掘技术建立更深入的数据剖析，并架构精准的预测模式，呈现真正智能型、 个人化的网络服务是Web 挖掘的方向。

**12.1.5** **数据挖掘的经典算法**

数据挖掘的经典算法如下所述。

● C4.5: 是机器学习算法中的一种分类决策树算法，其核心算法是 ID3 算法。

● K-means: 是一种聚类算法。

● SVM: 一种监督式学习的方法，广泛运用于统计分类以及回归分析中。

● Apriori: 是一种最有影响的挖掘布尔关联规则频繁项集的算法。

● EM: 最大期望值法。

● pagerank: 是 Google 算法的重要内容。

● Adaboost: 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器， 然后把弱分类器集合起来，构成一个更强的最终分类器。

● KNN: 是一个理论上比较成熟的方法，也是最简单的机器学习方法之一。

● Naive Bayes: 在众多分类方法中，应用最广泛的有决策树模型和朴素贝叶斯模型。

● Cart: 分类与回归树，在分类树下面有两个关键的思想，第一个是关于递归地划分 自变量空间，第二个是用验证数据进行剪枝。

**1.Apriori 算法集**

Apriori 算法是使用候选项集找频繁项算法，是一种最有影响的挖掘布尔关联规则频繁 项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单 层、布尔关联规则。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

该算法的基本思想是：首先找出所有的频集，这些项集出现的频繁性至少和预定义的 最小支持度一样。然后由频集产生强关联规则，这些规则必须满足最小支持度和最小可信 度。然后使用第1步找到的频集产生期望的规则，产生只包含集合的项的所有规则，其中 每一条规则的右部只有一项，这里采用的是中规则的定义。 一旦这些规则被生成，那么只 有那些大于用户给定的最小可信度的规则才被留下来。为了生成所有频集，使用了递推的 方法。可能产生大量的候选集，以及可能需要重复扫描数据库是 Apriori 算法的两大缺点。





**大数据基础与应用**

**2.** **划分算法**

划分算法的执行过程是先把数据库从逻辑上分成几个互不相交的块，每次单独考虑一 个分块并对它生成所有的频集，然后把产生的频集合并，用来生成所有可能的频集，最后 计算这些项集的支持度。这里分块的大小选择要使得每个分块可以被放入主存，每个阶段 只需被扫描一次。而算法的正确性是由每一个可能的频集至少在某一个分块中是频集保证 的。该算法是可以高度并行的，可以把每一分块分别分配给某一个处理器生成频集。产生 频集的每一个循环结束后，处理器之间进行通信来产生全局的候选 k-项集。通常这里的通 信过程是算法执行时间的主要瓶颈；而另一方面，每个独立的处理器生成频集的时间也是 一个瓶颈。

**3.FP—** **树频集算法**

FP—树频集算法不产生候选挖掘频繁项集的方法，而是采用分而治之的策略，在经过 第一遍扫描之后，把数据库中的频集压缩进一棵频繁模式树，同时依然保留其中的关联信 息，随后再将频繁模式树分化成一些条件库，每个库和一个长度为1 的频集相关，然后再 对这些条件库分别进行挖掘。当原始数据量很大的时候，也可以结合划分的方法，使得一 个频繁模式树可以放入主存中。实验表明，频繁树生长对不同长度的规则都有很好的适应 性，同时在效率上较 Apriori 算法有巨大的提高。

**12.2** **大数据挖掘技术**

准确而高效地从丰富的大数据中筛选出有价值的信息已经成为迫切需要，于是大数据 挖掘技术应运而生，并显示出强大的解决问题的能力，已成为未来重要的技术之一。

大数据挖掘是针对大数据的计算机辅助分析，是智能化处理模式。这是一种新的探索 领域，过去由于计算机联网水平、处理水平、信息积累能力的限制，人们看不到大数据这 个领域，因此就无法针对大数据进行建模处理。由于科技发展，出现了大数据的生态环境。 现在大数据呈现在人们眼前，人们有了新的视野，这时急需的是大数据理论和大数据挖掘 的方法以及数据运用。

积累的数据越多，所蕴含的价值也越大，需要一种方法把这些价值挖掘出来。从浩如 烟海的数据中寻觅有价值的信息的同时，用户还需要保存已经使用过的数据。因为这些旧 数据可以与未来收集的新数据进行对比，依然具有潜在的利用可能。与以往相比，除了有 能力存储更多的数据量之外，还需要面对更多的数据类型。这些数据的来源包括网上交易、 网络社交活动、自动传感器、移动设备以及科学仪器等。除了固定的数据生产源，各种交 易行为还可能加快数据的积累速度。例如，社交类多媒体数据的爆炸性增长来源于网上交 易和记录行为。数据永远都在增长之中。





第12章 大数据挖掘

**12.2.1** **大数据挖掘关键技术**

在传统的数据挖掘中，采用了随机抽样方法来获取小样本。大数据的出现克服了数据 挖掘中的数据不足，用大样本和全样本代替小样本，克服了样本数量不足的弊端，基于大 数据的挖掘技术可以使以前不能解决的问题变得可以解决。

对于大数据来说，即使计算机能够高效地处理它们，但大量无用的垃圾数据，只能给 系统带来负担，并增加存储设备和主机等设备成本。这就需要数据处理过程中，根据特定 的规则和参数，对涌进的数据流进行清洗和分析，并自动决策选择处理数据，不需要人工 干预。大数据处理平台提供了一系列针对大数据的数据分析和挖掘技术，例如数据清洗、 数据分析、数据挖掘和数据可视化等技术用来支持各领域的大数据处理及应用。数据挖掘 已在各领域广泛应用，只要该产业拥有分析价值与需求的数据仓储或数据库，都可以利用 挖掘工具进行数据挖掘。

大数据挖掘的目的是从大数据中挖掘出未知且有用的知识。只有通过大数据挖掘，大 数据的价值才得以体现，所以挖掘对大数据的利用具有举足轻重的意义。大数据挖掘有两 个重要的基本问题，即挖什么与怎么挖。挖什么决定了从数据中抽取什么样的信息、统计 什么样的规律，是数据的收集、处理和挖掘的全过程中需要考虑的问题；怎么挖决定怎样 具体进行抽取与统计，仅限于挖掘本身。怎么挖是数据挖掘技术研究的核心。因为挖什么 决定了挖掘结果的价值，所以挖什么在数据挖掘的应用中更为重要。

大数据挖掘关键技术主要包括大数据存储技术、云计算技术、并行技术、面向数据挖 掘的隐私保护技术和数据挖掘集成技术。

**1.** **大数据存储技术**

大数据存储技术主要包括并行存储体系架构、高性能对象存储技术、并行 I/O 访问技 术、存储系统高可用技术、嵌入式存储操作系统、数据保护与安全体系、绿色存储等。

数据存储系统为云计算、物联网等新一代高新技术产业提供核心的存储基础设施，为 一系列重大工程等起到了核心支撑和保障作用，存储系统已经使用到石油、气象、金融、 电信等重要行业与部门。发展具有自主知识产权、达到国际先进水平的数据存储系统可以 满足许多重大行业快速增长的大数据存储需要，并创造出巨大的经济效益。

**2.** **云计算技术**

云计算的相关应用主要包括云物联、云安全和云存储。云存储是在云计算概念上延伸 和发展出来的新概念，云存储通过集群应用、网格技术或分布式文件系统等功能，将网络 中大量的、不同类型的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储 和业务访问功能的一个系统。

当云计算系统运算和处理的核心是大量数据的存储和管理时，云计算系统就需要配置 大量的存储设备，那么云计算系统就转变成为一个云存储系统，所以云存储是一个以数据 存储和管理为核心的云计算系统。



**大数据基础与应用**

**3.** **并行技术**

当数据挖掘的对象是一个大数据集或许多广泛分布的数据源时，效率就成为数据挖掘 的瓶颈。随着并行处理技术的快速发展，用并行处理的方法来提高数据挖掘效率的需求越 来越大。并行数据挖掘涉及的体系结构和算法方面的技术如下所述。

① 硬件平台的选择(共享内存的或分布式的)。

② 并行的策略(任务并行、数据并行或任务并行与数据并行结合)。

③ 负载平衡的策略(静态负载平衡或动态负载平衡)。

④ 数据划分的方式(横向的或纵向的)等。

处理并行数据挖掘的策略主要涉及并行关联规则挖掘算法、并行聚类算法和并行分类 算法。

**4.** **面向数据挖掘的隐私保护技术**

应用数据挖掘技术，在产生价值的同时也随之出现了隐私泄露的问题。如何在防止隐 私泄露的前提下进行数据挖掘是现实迫切的技术需求。基于隐私保护的数据挖掘应用了数 据扰乱、数据重构、密码学等技术，能够在保证足够精度和准确度的前提下，使数据挖掘 者在不触及实际隐私数据的同时，仍能进行有效的挖掘工作。基于隐私保护的数据挖掘技 术可从以下四个层面进行分类。

① 基于数据的分布情况，可以分为原始数据集中式和分布式两大类隐私保护技术。

② 基于原始数据的隐藏情况，可以分为对原始数据进行扰动、替换和匿名隐藏等隐私 保护技术。

③ 基于数据挖掘技术，可以分为针对分类挖掘、聚类挖掘、关联规则挖掘等隐私保护 技术。

④ 基于隐藏内容，可以分为原始数据隐藏、模式隐藏等。

**5.** **数据挖掘集成技术**

数据挖掘体系框架由数据准备体系、建模与挖掘体系、结果解释与评价体系三部分组 成，其中核心的部分是建模与挖掘体系，它主要是根据挖掘主题和目标，通过挖掘算法和 相关技术，如统计学、人工智能、数据库、相关软件技术等，对数据进行分析，挖掘出数 据之间内在的联系和潜在的规律。数据挖掘的集成可分为下述几类。

① 数据挖掘算法的集成。

② 数据挖掘与数据库的集成。

③ 数据挖掘与数据仓库的集成。

④ 数据挖掘与相关软件技术的集成。

⑤ 数据挖掘与人工智能技术的集成等。

**12.2.2** **大数据挖掘策略**

数据挖掘过程的第一步是采集数据，被采集的数据的质量直接影响了能从数据中挖掘





**第12章** **大数据挖掘**

出什么样的知识。没有某方面的数据，也就无法从中进行相关的挖掘。但是，由于存储与 处理大数据代价巨大，提高数据挖掘效率的关键在于仅采集对记录有用的数据。所以要对 收集到的数据内容进行合理的判断，将可能有用的数据全部采集与记录下来。要做到这一 点就必须对挖掘场景做出合理的设想。场景是多种意图的组合，只有对数据挖掘的内容有 比较清晰的想法，才能对数据的收集范围有比较明确的界定，所以挖掘场景设计必不可少。

由于很难事先穷尽所有可能的挖掘场景，所以也就很难完全准确地判断应该收集哪些 数据。基于上述考虑，在存储、处理能力允许的条件下，尽量从多方面收集数据。显然， 更多的数据可以对数据挖掘产生帮助。

将相关数据整合在一起用于挖掘可以让数据发挥更大作用。数据整合有助于了解事物 的全貌，发现未知的关系，挖掘深层的知识，提升预测的准确率。决定从数据中挖掘什么, 首先需要对数据进行认真细致的观察，从而对数据有深入的了解。只有对数据有深刻的认 识，才有可能从中挖掘出深层的知识。

**1.** **关联规则发现**

(1)关联规则定义

数据关联是数据库中存在的一类重要的可被发现的知识。如果两个或多个变量的取值 之间存在某种规律性就称之为关联。关联分析的目的是找出数据库中隐藏的关联网。有时 并不知道数据仓库中数据的关联函数，即使知道也是不确定的，因此关联分析生成的规则 带有可信度。关联规则挖掘发现，大量数据中项集之间有趣的关联或相关联系。在数据挖 掘中，关联规则是一个重要的概念与技术，主要用于发现隐藏在数据仓库中有意义的联系。 例如超市为了能够准确了解顾客在其门店的购买习惯，对顾客的购物行为进行购物篮分析， 想知道顾客经常购买的商品。数据仓库里集中了其各门店的详细原始交易数据。在这些原 始交易数据的基础上，利用数据挖掘方法对数据进行分析和挖掘。

(2)关联规则分类

按照不同情况，关联规则可以进行如下分类。

① 基于处理的变量类别分类。基于处理的变量的类别，可以分为布尔型关联规则和数 值型关联规则。布尔型关联规则处理的值都是离散的种类变量，而数值型关联规则可以和 多维关联或多层关联规则结合起来，对数值型字段进行处理。数值型关联规则中也可以 包含种类变量。例如，性别=“女”=>职业=“秘书”规则，是布尔型关联规则；性别=“女” =>avg (收入)=2300 规则，涉及的收入是数值类型，所以是一个数值型关联规则。

② 基于数据的抽象层次分类。基于数据的抽象层次，可以分为单层关联规则和多层关 联规则。在单层的关联规则中，所有的变量没有考虑到现实的数据具有层次性，而在多层 的关联规则中，对数据的多层性已经进行了充分的考虑。

例 如 ，IBM 台式机=>Sony 打印机，是一个细节数据上的单层关联规则；台式机=>Sony 打印机，是一个较高层次和细节层次之间的多层关联规则。

③ 基于数据的维数分类。基于规则中涉及的数据的维数，关联规则可以分为单维的和 多维的。在单维的关联规则中，只涉及数据的一个维，如用户购买的物品；而在多维的关 联规则中，要处理的数据将涉及多个维。也就是说，单维关联规则是处理单个属性中的一





**大数据基础与应用**

些关系，多维关联规则是处理各个属性之间的某些关系。例如，啤酒=>尿布，这条规则只 涉及用户购买的物品；性别=“女”=>职业=“秘书”,这条规则就涉及两个字段的信息， 是两个维上的一条关联规则。

(3)关联规则的挖掘过程

关联规则挖掘过程主要包含下述两个阶段。

① 从资料集合中找出所有的高频项目组。高频是指某一项目组出现的频率相对于所有 记录高而言，必须达到某一水平。 一个项目组出现的频率称为支持度，以一个包含A 与 B 两个项目的项目组为例，如果支持度大于等于所设定的最小支持度阈值，则{A,B} 称为高 频项目组。关联规则挖掘的第一阶段必须从原始资料集合中找出所有高频项目组。

② 由高频项目组中产生关联规则。从高频项目组产生关联规则是指，利用前一步骤的 高频 k 项目组来产生规则，在最小信赖度的条件下，如果一规则所求得的信赖度满足最小 可信度，则称此规则为关联规则。例如，经由高频k 项目组{A,B} 所产生的规则AB, 如果 信赖度大于等于最小可信度，则称 AB 为关联规则。关联规则挖掘的第二个阶段是要产生 关联规则。

例如，使用关联规则挖掘技术对交易资料库中的记录进行资料挖掘，过程如下：

a. 首先设定最小支持度与最小可信度两个阈值，即最小支持度 min\_support=5% 且最 小信赖度 min\_confidence=70%。

b. 符合该超市需求的关联规则必须同时满足以上两个条件。如果经过挖掘过程所找到 的关联规则满足条件，将可接受「A,B」 的关联规则描述为

Support(A,B)>=5%∩Confidence(A,B)>=70%

此应用范例中的意义为：在所有的交易记录资料中，至少有5%的交易呈现A 与 B 这 两项商品被同时购买的交易行为。Confidence(A,B)>=70% 的意义为：在所有包含A 的交易 记录资料中，至少有70%的交易会同时购买B。

c. 如果有某消费者出现购买A 的行为，超市将可推荐该消费者同时购买B。这个商品 推荐的行为则是根据「A,B」 关联规则，支持了大部分购买A 的交易，会同时购买B 的消 费行为。

从上述的介绍中可以看出，关联规则挖掘通常适用于记录中的指标取离散值的情况。 如果原始数据库中的指标值取连续的数据，则在关联规则挖掘之前应该进行数据离散化。

**2.** **相似项发现**

从数据中发现相似项是数据挖掘的基本问题，相似项发现的典型应用是Web 网页的近 似查重，进而判断是否是抄袭的网页。相似项的寻找可以归为寻找具有相对较大交集的集 合问题，但也可以应用在任意空间下的距离度量理论来定义相似度问题。

利用shigling 技术将文本相似问题转换为集合运算问题。

利用最小哈希技术对大集合压缩，并且可以基于压缩后的结果推到原始集合的相似度。

任意类型的相似项搜索中存在的一个问题是即使对各项之间的相似度计算非常简单， 但是由于项目过多，无法对所有项对来检测相似度。为此，提出了局部敏感哈希技术，其





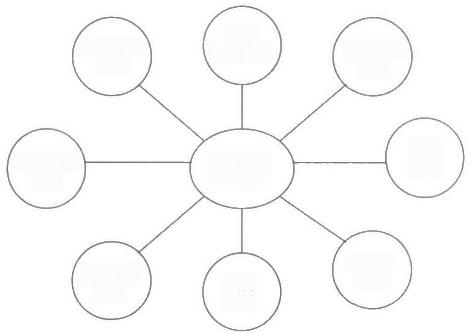
**第12章** **大数据挖掘**

主要思想是将搜索方位集中在那些可能相似的项对。

在这里仅介绍了关联规则发现和相似项发现问题，其他的聚类方法、数据流挖掘、链 接分析及推荐技术在本书的有关章节已简单介绍。

**12.3** **大数据挖掘应用**

大数据挖掘已在零售业、制造业、财务金融保险、通信业以及医疗服务等许多领域得 以应用。如图12-7所示，在销售数据中挖掘顾客的消费习性、可由交易记录找出顾客偏好 的产品组合、找出流失顾客的特征与推出新产品的时机点等都是零售业常见的实例。在应 用数据挖掘技术之后，促进了直效行销的发展，例如利用数据挖掘来分析顾客群的消费行 为与交易记录。能够结合基本数据，并依其对品牌价值等级的高低来区隔顾客，进而达到 差异化行销的目的。制造业对数据挖掘的需求多运用在品质管控方面，由制造过程中找出 影响产品品质最重要的因素，以期提高作业流程的效率。



物流 解决方案

数据智能 挖掘系统

品质

管理

市场行为 分析

物流需求 分析

信贷制度 审查

股票投资 战略

营业额 预测

客户 分类

图12-7 大数据挖掘的应用

电话公司、信用卡公司、保险公司以及股票交易行业每年因为欺诈行为而造成的损失 非常大，因此，十分重视欺诈行为的侦测，数据挖掘可以从一些信用不良的客户数据中找 出相似特征并预测可能的欺诈交易，达到减少损失的目的。财务金融业可以利用数据挖掘 来分析市场动向，并预测个别公司的营运以及股价走向。数据挖掘的另一个独特的用法是 在医疗业，用来预测手术、用药、诊断和流程控制的效率。

**12.3.1** **市场营销**

在市场营销过程中，给特定的人群发送一些消息，说明可以提供优惠，促进用户购买 某些产品，进而确定哪些人会更倾向于接受信息等，主要考虑如下四个因素。





**大数据基础与应用**

**1.** **用户的属性**

用户的属性包括性别、年龄、职业、公司、行业、地点等，这些相关信息属于静态的 信息，也是所有行业的一种基本数据信息。

**2.** **用户的行为数据**

以下例说明用户的行为数据的作用。某用户决定去买一个扫描仪，在此之前将做很多 的事情，最后才会决定购买。如果只是看用户买它的一条记录，那么它只是一条数据记录， 但是之前为什么决定买这个扫描仪，可能会有几千条、几万条不同的信息记录下来。这些 信息表明了用户的一系列行为，才导致最后的购买决定。对于这个问题，数据量要大很多， 从分析量上来看，难度也高很多，但是增加了预测的准确性。

**3.** **与社交网络相关的信息**

与社交网络相关的信息将对预测造成非常大的影响。相似的人往往做出接近的决定。 物以类聚，人以群分，在社交网上相连接的人往往具有共同的爱好和兴趣。

上述三个因素的数据组成了需要考察的数据。当然，每个因素中又有很多的数据。实 际上，为了提高预测的准确度，使用了很多的精力去采集这些数据，然后再不断地扩充。 这是由于随着数据量的增加，预测的准确度也会越来越高。

**4.** **时间序列**

除了上述因素之外，时间序列也非常重要。很多时候做的这件事情离现在有多久，会 对预测产生很大影响。例如，购买者在网站上看了一个扫描仪相关的描述。如果第二天销 售者就跟购买者说，会给50%的折扣，那购买者买下的可能性是很大的，因为购买者已经 有了很大的兴趣。如果一个月之后，销售者再来问购买者，那么购买者可能就没有很大的 兴趣了。这只是一个方面，但在整个时间序列中，多久之前发生这件事情很重要。

另外，在整个时间序列中，行为的变化也很重要。例如，购买者看这个扫描仪的介绍， 如果以前是一个月看一次，但最近变成一个礼拜看一次，或者说变成一天看一次，整个时 间序列就能说明购买者的关注点不断增加。所以时间序列中的这种模式也非常重要。

上述四个因素需要大量的数据进行挖掘，才能达到所需要的结果。

**12.3.2** **销售**

在某超市的一个角落，X 品牌的销售品摆放在某处。来自X 品牌公司的业务员每天例 行公事地来到这个点，拍摄10张照片：怎么摆放销售品、位置有什么变化、高度如何等， 这样的点每个业务员一天要跑15个，按照规定，下班之前150张照片就被传回了公司总部。 每个业务员，每天产生的数据量在10MB, 这似乎并不是个大数字。但公司在全国有10000 个业务员，这样每天的数据就是100GB, 每月为3TB 。当这些图片如同雪片一样进入公司 总部的机房时，这家公司的CIO 就有这么一种感觉：守着一座金山，却不知道从哪里挖下 第一锹。





**第12章** **大数据挖掘**

CIO 想知道的主要问题是：怎样摆放销售品更能促进销售，什么年龄的消费者在销售 品堆前停留更久，他们一次购买的量多大，气温的变化让购买行为发生了哪些改变，竞争 对手的新包装对销售产生了怎样的影响……不少问题目前也可以回答，但它们更多是基于 经验，而不是基于数据。

从2008年开始，业务员拍摄的照片就这么被收集起来，如果按照数据的属性来分类， 图片属于典型的非结构化数据。要系统地对非结构化数据进行分析是 CIO 的下一步计划。 如果超市、金融公司与销售品有某种渠道来分享信息，如果类似图像、视频和音频资料可 以系统分析，如果人的位置有更多的方式可以被监测到，那么摊开在CIO 面前的就是一幅 基于人消费行为的画卷，而描绘画卷的是一组组复杂的数据。

在日常运营中，产生了销售、市场费用、物流、生产、财务等数据，这些数据都是通 过工具定时抽取后展现。这个展现的过程长达24小时，也就是说，在24小时后，物流、 资金流和信息流才能汇聚到一起，彼此关联形成一份有价值的统计报告。当公司的每月数 据积累达到3TB 时，这样的速度导致公司每个月财务结算都要推迟一天。更重要的是，公 司的决策者们只能依靠数据来验证以往的决策是否正确，或者对已出现的问题做出纠正， 仍旧无法预测未来。当公司采用了基于大数据的创新平台之后，同等数据量的计算速度从 过去的24小时缩短到了0.67s, 几乎可以做到实时计算结果，这使很多不可能的事情变为 了可能。

**12.3.3** **物流**

某贸易有限公司与制鞋公司合作已有13年，拥有100多家制鞋公司门店。但是，库存 积压问题很严重，但通过合作解决问题，生意再次回到了正轨。在最初降价、打折等清库 存的应急措施结束后，基于外部环境、消费者调研和门店销售数据的收集、分析，制鞋公 司和经销商们走上了正轨。

经销商每天都会收集门店的销售数据，并将它们上传至制鞋公司。收到数据后，制鞋 公司对数据做整合、分析，再用于指导经销商卖货。研究这些数据，让制鞋公司和经销商 们可以更准确地了解当地消费者对商品颜色、款式、功能的偏好，同时知道什么价位的产 品更容易被接受。制鞋公司产品线丰富，过去，面对展厅里各式各样的产品，经销商很容 易按个人偏好下订单。现在，制鞋公司用数据说话，帮助经销商选择最适合的产品。首先， 从宏观上看， 一、二线城市的消费者对品牌和时尚更为敏感，可以重点投放采用前沿科技 的产品、运动经典系列的服装以及设计师合作产品系列。在中小城市，消费者更关注产品 的价值与功能。如纯棉制品这样高性价比的产品，在这些市场将更受欢迎。其次，制鞋公 司会参照经销商的终端数据，给予更具体的产品订购建议。例如，制鞋公司可能会告诉某 终端市场的经销商，在其辖区，普通跑步鞋比添加了减震设备的跑鞋更好卖。至于颜色， 比起红色，当地消费者更偏爱蓝色。推动这种订货方式，制鞋公司得到了经销商们的认可。 这样，将来就会大大减少库存问题。

大数据挖掘使制鞋公司有了许多新的发现。同为一线城市，地理位置不同，则消费趋 势不同。还有， 一线城市消费者的消费品位和习惯更为成熟，当地消费者需要不同的服装





**大数据基础与应用**

以应对不同场合的需要，上班、吃饭、喝咖啡、去夜店，需要不同风格的多套衣服。但在 中小城市， 一位女性往往只要有应对上班、休闲、宴请三种不同风格的服饰就可以。两相 对比， 一线城市显然为制鞋公司提供了更多细分市场的选择。

对大数据的运用，也顺应了制鞋公司战略转型的需要。库存危机后，制鞋公司从批发 型公司转为零售驱动型公司，它过去只关注把产品卖给经销商，现在变成了将产品卖到终 端消费者手中的有力推动者。而数据收集分析，恰恰能让其更好地帮助经销商提高销售率。

制鞋公司与经销商伙伴展开了更加紧密的合作，以统计到更为确切可靠的终端消费数 据，有效帮助制鞋公司重新定义了产品供给组合，从而在适当的时机，将符合消费者品位 的产品投放到相应的区域市场。 一方面降低了库存，另一方面增加了单店销售率。卖得更 多，获取更高的利润。

**12.3.4** **CRM**

在蓬勃发展的市场环境中，大数据所带来的机遇前所未有，这将是中国市场的营销者 们预期取得大回报的最佳时机。这也正是 CRM 服务商发展的良机，以数据为本、分析为 先的CRM, 大数据摆在面前势必不能放过。大数据必将成为CRM 产业的催化剂，使之成 为多元化大数据时代营销者的全新工具。

不同行业、不同环境下的 CRM 应用差异很大，应用到的具体数据挖掘技术与方法也 将不同。虽然数据挖掘技术和方法层出不穷，但也难以涵盖全部的技术和方法。不同的 CRM 应用到的数据挖掘技术很多，也很复杂，但是主要用在客户细分、客户响应分析、提升客 户价值和客户流失分析等方面。

**1.** **客户细分**

客户细分是指根据客户的性别、收入、交易行为特征等属性将其细分为具有不同需求 和交易习惯的群体，同一群体中的客户对产品的需求及交易心理等方面具有相似性，而不 同群体间差异较大。客户群体细分可以使企业在市场营销中制定正确的营销策略，通过对 不同类别客户提供有针对性的产品和服务，提高客户对企业和产品的满意度，以获取更大 的利润。

客户细分可以采用分类的方法，也可以采用聚类的方法。比如，可以将客户分为高价 值和低价值的客户，然后确定对分类有影响的因素，再将拥有相关属性的客户数据提取出 来，选择合适的算法对数据进行处理得到分类规则。使用聚类的方法，则在之前并不知道 客户可以分为几类，在将数据聚类后，再对结果数据进行分析，归纳出相似性和共性。

每一类别的客户具有相似性的属性，而不同类别客户的属性也不同，从而确定特定消 费群体或个体的兴趣、消费习惯、消费倾向和消费需求，进而推断出相应消费群体或个体 下一步的消费行为。细分可以让用户从比较高的层次上来查看整个数据库中的数据，也使 得企业可以针对不同的客户群采取不同的营销策略，有效地利用有限的资源。合理的客户 细分是实施客户关系管理的基础。





**第12章** **大数据挖掘**

**2.** **客户响应分析**

在大多数商业领域中，业务发展的主要指标里都包括新客户的获取能力。新客户的获 取包括发现那些对产品不了解的顾客，但他们可能是公司产品的潜在消费者，也可能是以 前接受公司的竞争对手服务的顾客。在寻找新客户之前，企业应该确定哪些客户是可能的 潜在客户、哪些客户容易获得、哪些客户较难获得，从而使企业有限的营销资源得到最合 理的利用。因此，预测潜在客户对企业销售推广活动的反应情况是获得客户的前提，由于 潜在客户的数量日益庞大，如何提高市场促销活动的针对性和效果成为获取新客户的关键 问题。数据挖掘可以帮助企业识别出潜在的客户群，提高客户对市场营销活动的响应率， 使企业做到心中有数、有的放矢。根据企业给定的一系列客户资料及其他输入，数据挖掘 工具可以建立客户反应预测模型，利用这个模型可以计算出客户对某个营销活动的反应指 标。企业根据这些指标就可以找出对企业所提供的服务感兴趣的客户，进而达到获取客户 的目的。数据挖掘技术中的关联分析、聚类和分类功能可以很好地完成这种分析。

**3.** **提升客户价值**

交叉销售是指企业向原有客户销售新的产品或服务的营销过程，它不仅是通过对现有 客户扩大销售来增加利润的一个有效手段，而且还是提升企业形象、培养客户忠诚度、保 障企业可持续发展的重要战略。

公司与其客户之间的商业关系是一种持续的、不断发展的关系。在客户与公司建立起 这种双向的商业关系之后，可以有很多种方法来优化这种关系，延长这种关系的时间。在 维持这种关系期间，增加互相的接触，努力在每一次互相接触中获得更多的利润。而交叉 销售就是这种工具，即向现有的客户提供新的产品和服务的过程。

在交叉销售活动中，数据挖掘可以帮助企业分析出最优的销售匹配方式。在企业所掌 握的客户信息，尤其是以前购买行为的信息中，可能正包含着这个客户决定他下一个购买 行为的关键因素，甚至决定因素。通过相关分析，数据挖掘可以帮助分析出最优的、最合 理的销售匹配。 一般过程是：首先，分析现有客户的购买行为和消费习惯数据，然后用数 据挖掘的一些算法对不同销售方式下的个体行为进行建模；其次，用建立的预测模型对客 户将来的消费行为进行预测分析，对每一种销售方式进行评价；最后，用建立的分析模型 对新的客户数据进行分析，以决定向客户提供哪一种交叉销售方式最合适。有几种数据挖 掘方法可以应用于交叉销售。关联规则分析，能够发现顾客倾向于关联购买哪些商品；聚 类分析，能够发现对特定产品感兴趣的用户群；神经网络、回归等方法，能够预测顾客购 买该新产品的可能性。

相关分析的结果可以用在交叉销售的两个方面： 一方面是对于购买频率较高的商品组 合，找出那些购买了组合中大部分商品的顾客，向他们推销遗漏的商品；另一方面是对每 个顾客找出比较适用的相关规律，向他们推销对应的商品系列。

**4.** **客户流失分析**

随着企业竞争越来越激烈，企业获取新客户的成本不断上升。对大多数企业而言，获 取一个新客户的花费大大超过保持一个已有客户的费用，保持原有客户的工作越来越有价



**大数据基础与应用**

值，这已经成为大多数企业的共识。保留一个客户的时间越长，收取在这个客户身上所花 的初期投资和获取费用的时间也越长，从客户身上获得的利润就越多。但由于各种因素的 不确定性和市场的不断增长，以及一些竞争对手的存在，很多客户为了寻求更低的费用和 其他服务商为新客户提供的更多的额外优惠条件，不断地转向另一个服务商。我们把客户 从一个服务商转向另一个服务商的行为称为客户转移。为了分析出是哪些主要因素导致客 户转移，并可以有针对性地挽留那些有离开倾向的客户，可以通过使用数据挖掘工具为已 经流失的客户建模，识别导致他们转移的模式，然后用这些找出当前客户中可能流失的客 户，以便企业针对客户的需要，采取相应的措施防止客户的流失，进而达到保持原有客户 的目的。

解决客户流失问题，首先需要明确流失的是什么样的客户。如果流失的是劣质客户， 企业求之不得；如果流失的是优质客户，企业则损失巨大。企业优质客户的稳定期越长， 企业与其维持关系的成本越低，获得的收益越大。因此，为保持优质客户，需要先辨识优 质客户。这通过前面的客户细分就可以完成，分析出客户盈利能力，辨识和预测客户的优 劣。当能够辨识出客户的优劣时，首先，根据已流失客户数据，利用决策树、神经网络等 进行分析挖掘，发现流失客户特征；然后，对现有客户消费行为进行分析，以确定每类客 户流失的可能性，其中着重于发现那些具有高风险转移可能性并具有较高商业价值的客户， 在这些客户转移到同行业其他服务商那里之前，采取相应的商业活动措施来保持住这些有 价值的客户。我们把这个过程叫作客户保留或客户保持。

在选择数据挖掘工具时，若希望能够对客户进行细分，并且能够对客户流失的原因有 比较清晰的了解，那么决策树工具是比较好的选择。尽管其他的一些数据挖掘技术，如神 经元网络也可以产生很好的预测模型，但是这些模型很难理解。当用这些模型做预测分析 时，很难对客户的流失原因有深入的了解，更得不到对付客户流失的任何线索。在这种情况 下，也可使用细分技术和聚类技术来得到深入的了解，但用这些技术生成预测模型就相对复 杂得多。 一般来说，在客户保持的应用领域，大多使用分类回归决策树来生成预测模型。

综上所述，数据挖掘在CRM 中有着广泛的应用，从某个角度可以说数据挖掘是CRM 的灵魂。通过运用数据挖掘的相关技术，发现数据中存在的关系与规则，为管理者提供重 要的决策参考，用来制定准确的市场策略。并且，通过销售和服务等部门与客户交流，争 取最优化地满足客户的需求，提高客户忠诚度和满意度，提升客户价值，提高企业收益， 达到企业与客户的双赢局面。正是这一点，使得 CRM 得到了很大成功。

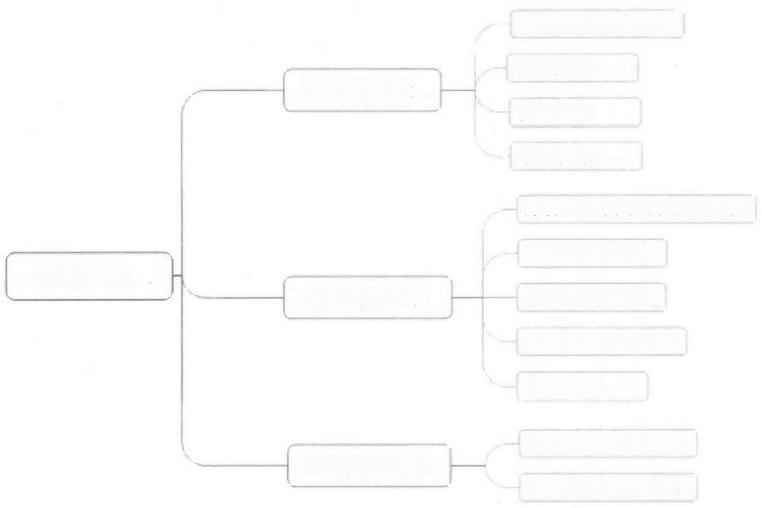
小结

数据挖掘不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联 系，从而促进信息的传递，实现规律与规则的发现。数据挖掘的基础技术现已经基本成熟， 并已经得到广泛应用。人们对大数据挖掘寄予厚望，大数据挖掘关键技术主要包括大数据 存储技术、云计算、并行技术、隐私保护技术和数据挖掘集成技术。大数据挖掘已在零售 业、制造业、财务金融保险、通信业以及医疗服务等许多领域得以应用。



**第13章** **大数据可视化**

本章主要内容



可视化技术的产生与发展

科学可视化

可视化技术概述

信息可视化

数据可视化

高可伸缩性的分布式并行可视化算法

并行图像合成算法

大数据可视化

并行颗粒跟踪算法

重要信息的提取与显示

原位可视化

大数据可视化分析概念

大数据可视化分析

大数据可视化分析方法

大数据科学可视化

在科学技术发现的长河中，人类的视觉发挥了杰出的作用。望远镜和显微镜等可视化 工具的出现，放大和扩展了人类的视觉功能，而新的可视化数据开发工具将会进一步拓展 人类的视野。

**13.1** **可视化技术概述**

人类的创造性不仅取决于人的逻辑思维，而且还取决于人类的形象思维，将数据映射 为视觉符号，充分利用人们的杰出视觉，帮助人们获取大数据中蕴含的信息。只有将大数 据可视化之后变成形象才能激发人的形象思维与想象力。除此之外，人类视觉还具有对大 量抽象的数据进行理解与分析的功能，也就是说，从表面上看来是无序的、杂乱无章的大 量数据中，找出隐藏其中的难以发现的规律，进而获得科学发现、工程开发、医疗诊断和 业务决策等的依据。从近年来的科学进展可以看出，可视化方面的新技术和关键技术对重 大科学发现起到重要作用。在大数据时代，大数据可视化技术研究不可避免，它的研究与

**大数据基础与应用**

发展将为科学的新发现创造新的手段和条件。

**13.1.1** **可视化技术的产生与发展**

数据可视化技术起源于20世纪50年代，首先出现了计算机图形学，利用计算机创建 出了首批图形图表。1987年，布鲁斯·麦考梅克、托马斯·德房蒂和玛克辛·布朗所撰写

*的美国国家科学基金会报告Visualization in Scientific Computing, 即《科学计算之中的可视*

化》,显著地促进了可视化领域的发展。在这份报告中强调了新的基于计算机的可视化技术 方法的必要性。随着计算机运算能力的迅速提升，建立了规模越来越大，复杂程度越来越 高的数值模型，从而构造了各种巨大的数据集。现在，不但利用医学扫描仪和显微镜之类 的数据采集设备产生大数据集，而且还利用各种存储方式的NoSQL 数据库来保存文本、日 志、网页和图像与声音等结构化和非结构化数据，因而，迫切需要新的计算机图形学技术 与方法来处理这些大数据并使之可视化。

**1.** **数据可视化技术的适用范围**

将科学计算之中的可视化称为科学可视化，科学可视化最初指的是作为科学计算之组 成部分的可视化，也就是科学与工程实践中对于计算机建模和模拟的运用。后来，数据的 种类增多和数量增大，包括那些来自商业、财务、行政管理、数字媒体等方面的非结构化 大数据。20世纪90年代初期，出现了信息可视化的研究领域，为许多应用领域之中对于 抽象的非结构性数据集的分析提供了支持。目前将科学可视化与信息可视化都归为数据可 视化。

由于数据可视化的范围不断地扩大，导致数据可视化成为一个处于不断发展与动态变 化的概念。数据可视化技术是较为高级的技术方法，这些技术方法允许利用图形、图像处 理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示， 对数据加以可视化解释，表明数据可视化所涵盖的技术方法更为广泛。

**2.** **信息展现内容**

实现信息展现是数据可视化的重要研究问题与手段。从数据可视化的基本信息展现角 度来看，展现的信息由统计图形和主题图两个主要的组成部分构成，内容包括：思维导图、 新闻显示、数据显示、链接显示、网站显示、文章与资源、工具与服务。

**3.** **信息展现方式**

上述的所有这些主题的特点全都与图形设计和列表密切相关。

(1)列表

将实验数据按一定规律用列表方式表达出来是记录和处理实验数据最常用的方法。表 格的设计要求对应关系清楚、简单明了、有利于发现相关量之间的物理关系。此外还要求 在标题栏中注明物理量名称、符号、数量级和单位等。根据需要还可以列出除原始数据以 外的计算栏目和统计栏目等。最后还要求写明表格名称、主要测量仪器的型号、量程和准 确度等级、有关环境条件参数，如温度、湿度等。



第13章 大数据可视化

(2)作图

作图法可以显式地表达物理量间的变化关系。从图线上还可以简便求出实验需要的某 些结果，如直线的斜率和截距值等，读出没有进行观测的对应点(内插法),或在一定条件 下从图线的延伸部分读到测量范围以外的对应点(外推法)。此外，还可以把某些复杂的函 数关系，通过一定的变换用直线图表示出来。

(3)常用的图表方法

● 柏拉图(排列图)

排列图是分析和寻找影响质量主要因素的一种工具，其形式用双直角坐标图，左边纵 坐标表示频数(如件数、金额等),右边纵坐标表示频率(如百分比表示)。分折线表示累 积频率，横坐标表示影响质量的各项因素，按影响程度的大小(即出现频数多少)从左向 右排列。通过对排列图的观察分析可抓住影响质量的主要因素。

● 直方图

直方图将一个变量的不同等级的相对频数用矩形块标绘的图表(每一矩形的面积对应 于频数),直方图又称柱状图、质量分布图，是一种统计报告图，由一系列高度不等的纵向 条纹或线段表示数据分布的情况。 一般用横轴表示数据类型，纵轴表示分布情况。

● 散点图

散点图表示因变量随自变量而变化的大致趋势，据此可以选择合适的函数对数据点进 行拟合。用两组数据构成多个坐标点，考察坐标点的分布，判断两变量之间是否存在某种 关联或总结坐标点的分布模式。

● 鱼骨图

鱼骨图是一种发现问题根本原因的方法，它也可以称为因果图。其特点是简捷实用， 深入直观。它看上去有些像鱼骨，问题或缺陷(即后果)标在鱼头之外。

FMEA

●

FMEA 是一种可靠性设计的重要方法。它是FMA (故障模式分析)和FEA (故障影响 分析)的组合。它对各种可能的风险进行评价、分析，以便在现有技术的基础上消除这些 风险或将这些风险减小到可接受的水平。

可视化分析是一个新的学科方向，可视化分析科学是指通过交互可视界面来进行分析、 推理和决策的科学。可视化分析与各个领域的数据形态、大小及其应用密切相关。

**4.** **思维导图**

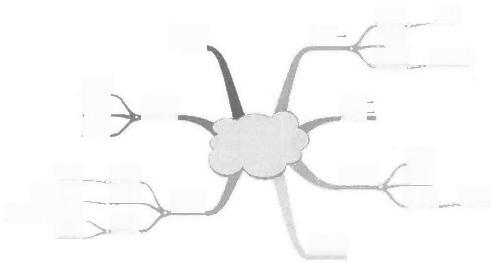
思维导图由英国人托尼·博赞 (Tony Buzan) 提出，是表达放射性思维的图形思维的有 力工具，简单而有效，是一种新型的思维工具。思维导图运用图文并茂的形式，把各级主 题的关系用相互隶属与相关的层级图表现出来，把主题关键词与图像、颜色等建立记忆链 接，思维导图充分运用左脑或右脑的机能，利用记忆、阅读、思维的规律，协助人们在逻 辑与想象之间平衡发展，从而开启了人类大脑的无限潜能，思维导图具有人类思维的强大 功能。

一个有关应用领域的思维导图如图13-1所示。





大数据基础与应用



上课

周五

听讲座

|  |  |
| --- | --- |
| 周二  周三  周四 | 上午  下午  晚上。 \_班会 |

补觉

出去购物

做下一周计划 休息

上午

下午

晚上

上午 下午 晚上

上午 下午 晚上

做计划

周六

周旦

周一

图13-1 思维导图

思维导图是一种将放射性思考具体化的方法。放射性思考是人类大脑的自然思考方式， 每一种进入大脑的资料，不论是感觉、记忆或想法，包括文字、数字、符号、食物、线条、 颜色、意象、节奏、音符等都可以成为一个思考中心，并由此中心向外发散出成千上万的 关节点，每一个关节点代表与中心主题的一个联结，而每一个联结又可以成为另一个中心 主题，再向外发散出成千上万的关节点，呈现出放射性立体结构，而这些关节点的联结可 以视为记忆，也就是个人数据库。

人类大脑惊人的存储能力使我们累积了大量的资料，经由思维导图的放射性思考方法， 使资料的存储、管理及应用更系统化而增加大脑运作的效率。同时，思维导图是最能善用 左脑或右脑的功能，利用颜色、图像、符码，不但可以协助记忆、增进创造力，也让思维 导图更轻松有趣，且具有个人特色及多面性。

思维导图以放射性思考模式为基础，收放自如，除了提供一个正确而快速的学习方法 与工具之外，将其运用在创意的联想与收敛、项目企划、问题解决与分析、会议管理等方 面，往往产生令人惊喜的效果。除了加速资料的累积量外，更多的是将数据依据彼此间的 关联性分层分类管理，它是一种展现个人智力潜能极至的方法，将可提升思考技巧，大幅 增进记忆力、组织力与创造力。它与传统笔记法和学习法具有量子跳跃式的区别，主要是 因为它源自脑神经生理的学习互动模式，并且开展人生而具有的放射性思考能力和多感官 学习特性。

思维导图为人类提供一个有效思维图形工具，运用图文并重的技巧，开启人类大脑的 无限潜能。心智图充分运用左脑或右脑的机能，协助人们在科学与艺术、逻辑与想象之间 平衡发展。思维导图完整的逻辑架构及全脑思考的方法得以广泛应用。

**5.** **信息展现涉及的领域**

从计算机科学的角度，可将这一领域划分为可视化算法与技术方法、立体可视化、信 息可视化、多分辨率方法、建模技术方法、交互技术方法与体系架构等子领域。

**6.** **数** **据** **与** **信** **息** **及** **知** **识**

数据、信息及知识是三个不同的概念。数据是符号的集合。信息是有用的数据，信息 不等同于知识，信息不能像知识那样去反映数据之间的内在联系。对于知识，可分成两类：





**第13章** **大数据可视化**

一类是无法用语言和文字来描述的知识，称之为隐知识；另一类是可以用语言和文字来描 述的知识，称之为显知识。由于信息可以用语言和文字来描述，所以信息是显知识。未来 在脑科学取得突破的基础上，将研制成功类人脑的生物计算机，从而开创人工智能的黄金 时代。但即使到那时，信息也不能完全表达人类全部的隐知识。只有将数据和信息用图形 和图像表示出来，才有可能为获得宝贵的隐知识创造条件。

数据可视化可以大大加快数据的处理速度，使产生的大量数据得到有效利用。可以在 人与数据、人与人之间实现图像通信，从而使人们能够观察到数据中隐含的现象。为发现 和理解科学规律提供有力工具，进而对计算和编程过程实现引导和控制，通过交互手段改 变过程所依据的条件，并观察其影响。可视化技术需要与数据处理和分析方法结合，提高 可视化的质量，同时给用户提供更完整的解决方案。

**7.** **交互式处理**

利用计算机进行科学计算和数据处理已有近60年的历史。长期以来，由于计算机技术 的限制，数据只能进行批处理而不能进行交互处理，也不能对计算过程进行干预和引导， 只能被动地等待计算结果的输出。而大量的输出数据也只能采用人工方式处理，或者使用 绘图仪输出二维图形。这种方法不仅不能及时地获得有关数据的直观、形象的整体概念， 而且还有可能丢失大量信息。近年来，来自超级计算机、天文、卫星、医学成像设备以及 地质勘探的数据与日俱增，使数据可视化成为迫切需要解决的问题。另一方面，由于计算 机的计算速度迅速提高，内存容量和磁盘空间不断扩大，网络功能日益增强，出现了使用 硬件来实现许多重要的图形生成及图像处理算法，运用数据可视化技术，可以直观、形象 地显示大数据和信息，并为进行交互处理的研究奠定了基础。

各种类型的信息源产生了大量数据，超出了人脑分析与解释这些数据的能力。主要是 缺乏大量数据的有效分析手段，浪费了大约95%的计算，这就严重阻碍了科学研究的进展。 在这种情况下，出现了可视化技术，并得到了迅速的发展。

数据可视化的成功应归于基本思想的完备性，依据数据及其内在模式和关系，利用计 算机生成的图像来获得深入的丰富知识，是利用人类感觉系统来操纵和解释错综复杂的过 程，涉及了不同学科领域的数据集以及大型抽象数据集合的模拟。

**13.1.2** **科学可视化**

可视化首先应用于科学与工程计算领域，并逐渐演化为广泛应用的可视化技术。

**1.** **科学可视化概念**

可视化技术是解释大量数据的最有效手段，因而首先在科学与工程计算领域中应用，

并发展为科学可视化的研究领域。应用可视化技术将数据转换成图形，给予人们意想不到 的洞察力，在各领域使科学家的研究方式发生了根本变化。可视化技术的应用广泛，大至 航天高速飞行模拟，小至分子结构的演示，无处不在。尤其在互联网时代，借助“互联网+” 技术能使远程可视化服务成为现实。可视区域网络的核心技术是可视化服务器硬件和软件。



**大数据基础与应用**

**2.** **科学可视化的过程**

科学可视化的主要过程是建模和渲染。

(1)建模

建立系统模型的过程又称模型化。建模是研究系统的重要手段和前提。凡是用模型描 述系统的因果关系或相互关系的过程都属于建模。因描述的关系各异，所以实现这一过程 的手段和方法多种多样。可以通过多种方法对系统本身运动规律进行分析，根据事物的机 理来建模，也可以通过对系统的实验或统计数据的处理，并根据关于系统的已有的知识和 经验来建模。在科学可视化的过程中，建模是指把数据映射成物体的几何图元。

(2)渲染

渲染是绘制真实感图形的主要技术，渲染能够把几何图元描绘成图形或图像。严格地 说，渲染就是根据基于光学原理的光照模型，计算物体可见面投影到观察者眼中的光亮度 大小和色彩的组成，并把它转换成适合图形显示设备的颜色值，从而确定投影画面上每一 像素的颜色和光照效果，最终生成具有真实感的图形。真实感图形是通过物体表面的颜色 和明暗色调来表现的，它与物体表面的材料性质、表面向视线方向辐射的光能有关，计算 复杂，计算量很大。渲染是三维计算机图形学中最重要的研究课题之一，并且在实际领域 与其他技术密切相关。

**13.1.3** **信息可视化**

自18世纪后期数据图形学诞生以来，抽象信息的视觉表达手段一直被人们用来揭示数 据及其他隐匿模式的奥秘。20世纪90年代期间问世的图形化界面，则使得人们能够直接 与可视化的信息之间进行交互，从而造就和带动了10多年来的信息可视化研究。信息可视 化通过利用人类的视觉能力来理解抽象信息的意思，从而加强人类的认知活动，使得达到 具有固定知觉能力的人类就能驾驭日益增多的数据。

20 世纪90年代以来才兴起的信息可视化，实际上源自其他领域。信息可视化已经在 图形学、视觉设计、计算机科学、人机交互，以及新近出现的心理学和商业方法等研究方 向上崭露头角。

**1.** **信息可视化概念**

信息可视化是跨学科领域的大规模非数值型信息资源的视觉展现，帮助人们理解和分 析数据。信息可视化中的交互方法允许用户与数据进行快速互动，以便更好地验证假设和 发现内在联系。信息可视化技术为人们提供了理解高维度、多层次、时空、动态、关系等 复杂数据的手段。与科学可视化相比，信息可视化则侧重于抽象数据集，如非结构化文本 或者高维空间中不具有固有的二维或三维几何结构的点的视觉展现。

信息可视化涉及了数据可视化、信息图形、知识可视化、科学可视化以及视觉设计方 面的所有进展。在这种层次上，如果加以充分适当的组织整理，任何事物都是一类信息： 表格、图形、地图，甚至包括文本在内，无论其是静态的还是动态的，都将为我们提供某 种方式或手段，从而让我们能够洞察其中的究竟，找出问题的答案，发现形形色色的关系，





**第13章** **大数据可视化**

或许还能让我们理解在其他形式的情况下不易发觉的事情。在科学技术研究领域，信息可 视化则一般适用于大规模非数字型信息资源的可视化表达。

信息可视化致力于创建以直观方式传达抽象信息的手段和方法。可视化的表达形式与 交互技术则是利用人类眼睛通往心灵深处的广阔带宽优势，使得用户能够看见、探索甚至 立即理解大量的信息。

各种各样数据结构的可视化需要新的用户界面以及可视化技术方法，现已发展成为一 门独立的学科，也就是信息可视化。信息可视化与经典的科学可视化是两个相关的领域， 但二者却有所不同。在信息可视化中，可视化的数据并不是某些数学模型的结果或者大型 数据集，而是具有自身内在固有结构的抽象数据。此类数据的例子包括：

● 编译器等各种程序的内部数据结构，或者大规模并行程序的踪迹信息；

● WWW 网站内容；

● 操作系统文件空间；

● 从各种数据库查询引擎那里所返回的数据，如数字图书馆。

信息可视化的另一项特点是所采用的工具侧重于宽广常用的环境，如普通工作站、 WWW 、PC 机等。这些信息可视化工具并不是为价格昂贵的专业化高端计算设备而定制。

信息可视化与可视化分析在目标和技术之间存在着部分重叠。虽然在这两个领域之间 还没有一个清晰的边界，但大致有三个方面可以加以区分。科学可视化主要处理具有地理 结构的数据，信息可视化主要处理树、图形等抽象式的数据结构，可视化分析则主要挖掘 数据背景的问题与原因。

通过数据可视化技术，可以发现大量金融、通信和商业数据中隐含的规律，从而为决 策提供依据。科学计算可视化技术是指空间数据的可视化技术，而信息可视化技术则是指 非空间数据的可视化技术。

**2.** **知识发现**

随着社会信息化的发展和网络的日益广泛的应用，信息源越来越庞大。除了需求对海 量数据进行存储、传输、检索及分类之外，更迫切需要了解数据之间的相互关系及变化趋 势。在激增的数据背后，隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析， 以便更好地利用这些数据。虽然数据库系统可以高效地实现数据的录入、查询、统计等功 能，但却无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的变化趋势。 另一方面，机器学习是用计算机模拟人类学习的一门科学，比较成熟的算法有神经网络、 遗传算法等。用数据库管理系统来存储数据，用机器学习的方法来分析和挖掘大量数据背 后的知识，这两者的结合促进了数据库中的知识发现 (Knowledge Discovery in Databases, KDD) 的产生。KDD 是一门交叉性学科，涉及机器学习、模式识别、统计学、智能数据库、 知识获取、数据可视化、高性能计算、专家系统等多个领域。KDD 可以应用在信息管理、 过程控制、查询优化、科学研究、决策支持和数据自身维护等许多方面。

KDD 的核心技术是数据挖掘，原始数据可以是结构化的，如关系型数据库中的数据， 也可以是半结构化的，如文本、图形数据，甚至是分布在网络上的非结构化数据。数据挖 掘的方法可以是基于数学的方法，也可以是演绎的或归纳的方法。通过数据挖掘可以发现



**大数据基础与应用**

多种类型的知识，包括反映同类事物共同性质的广义型知识、反映事物各方面特征的特征 型知识、反映不同事物之间属性差别的差异型知识、反映事物之间依赖或关联的关联型知 识、根据当前历史和当前数据推测未来数据的预测型知识、揭示事物偏离常规出现异常现 象的偏离型知识等。

**3.** **知识发现工具**

为了发现各种类型的知识。要采用多种发现知识的工具。为了使发现知识的过程和结 果易于理解和在发现知识过程中进行人机交互，要研究与发展发现知识的可视化方法。可 视化技术可以了解数据之间的相互关系及发展趋势。信息可视化不仅用图像来显示多维的 非空间数据，使用户加深对数据含义的理解，而且使用形象直观的图像来指导检索过程， 加快检索速度。在科学计算可视化中，显示的对象涉及标量、矢量及张量等不同类别的空 间数据，研究的重点放在如何真实、快速地显示三维数据场。而在信息可视化中，显示的 对象主要是多维的标量数据，通过选择显示方式设计和比较，便于用户了解多维数据及它 们相互之间的关系，其中更多地涉及心理学、人机交互技术等问题。

**4.** **信息可视化技术的应用**

信息可视化在商务、金融和通信等领域有着十分广阔的应用前景。目前正在开发更为 精细和高级的网络模型，以辅助将来的规划过程。但是另一方面，更复杂的发射和交换设 备，为现行网络的重构提供了更大的自由度和灵活性，造成在单个网络单元上运行的原始 数据不断增加。全部网络运行的最优化，需要有效地使用来自所有这些信号源，而且需要 在市场、网络规划和日常管理等传统的不同领域之间，进行信息的动态交换。覆盖物理网 络的是一个包括声音、数据和图像服务的广阔领域，其中每一项都有自己的数据和管理要 求。 此外，现代网络是一个覆盖很多国家和载体的国际性结构，因而其潜在的数据量和复 杂程度均以更大的数量级在递增。应用了信息可视化技术，通过测量大量运行参数，每天 要产生2GB 以上的数据。图形输出描绘了选择的运行参数的地理分布，以及所感兴趣的时 间间隔中的动画。每个区域中参数的最小值、最大值和平均值都可以用一个彩条图表示。 可视化在非空间数据中，如在财务指标或流通量统计中的应用，也引起了广泛的兴趣。很 多用于工程和科学应用中的可视化工具和技术能够很快地转移到财务和统计中来。可视化 应用成功的关键在于它具有为用户提供了交互式的研究数据和揭示那些用其他方法很困难 揭示的趋势、循环和模式的能力。在非空间数据范围内应用的一个典型例子是网络统计， 其中包括记录单个网络单元的特性、开关、较大区域或地理分组等。

城市景象可视化也是这方面的一项潜在的有用技术。城市景象是一个扩展的3D 条状 图，其中2D 域上的标量值表示为一个均匀网格上的街区或大楼。可视化表示出对一年中 的每个月，划分成十个地理带上呼叫失败率的统计资料。BT 已将城市景象应用于调查按月 按区的服务统计和传送系统运行性能。这些应用可以非常容易地用于金融信息，如每个区 域、每个时间段的股票收益特性，或按地理和按收入可视化挖掘，显示各个分行的货币流 通总量、总收入和现金运作统计的消费总量。

人工智能软件可以从异常现象中发现通过银行系统的非法活动。由于信息可视化对日





**第13章** **大数据可视化**

益显著的数据超载问题，可以提供近实时的解，它将对商务、金融和通信等领域的信息管 理，产生重要的影响。由此可见，数量日益增加的数据和信息是有用的，而关键在于尽快 从中提炼出有用的知识。

**13.1.4** **数据可视化**

**1.** **数据可视化概念**

数据可视化是关于数据的视觉表现形式的研究。数据可视化是指将大型数据集中的数 据以图形图像形式表示，数据可视化技术是指运用计算机图形学和图像处理技术，将数据 转换为图形或图像在屏幕上显示出来，并利用数据分析和开发工具发现其中未知信息的交 互处理的理论、方法和技术。数据可视化技术与计算机图形学、图像处理、计算机辅助设 计、计算机视觉及人机交互技术密切相关。数据可视化主要借助于图形化手段，清晰有效 地传达与沟通信息。

随着计算机技术的发展，数据可视化概念已大大扩展，不仅包括科学计算数据的可视 化，而且包括工程数据和测量数据的可视化。通常又把这种空间数据的可视化称为体视化 技术。近年来，随着数据仓库技术、网络技术、电子商务技术等的发展，可视化技术涵盖 了更广泛的内容，并进一步完善了数据可视化的概念。所谓数据可视化是对大型数据库或 数据仓库中的数据的可视化，它是可视化技术在非空间数据领域的应用，使人们不再局限 于通过关系数据表来观察和分析数据信息，还能以更直观的方式看到数据及其结构关系。 数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元元素表示，大量的数 据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度 观察数据，从而对数据进行更深入的观察和分析。

数据可视化技术包含以下几个基本概念。

① 数据空间：是由n 维属性和 m 个元素组成的数据集所构成的多维信息空间。

② 数据开发：指利用一定的算法和工具对数据进行定量的推演和计算。

③ 数据分析：指对多维数据进行切片、块、旋转等动作剖析数据，从而能多角度、多 侧面观察数据。

④ 数据可视化：是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和 开发工具发现其中未知信息的处理过程。

根据可视化的原理不同，可将现存的数据可视化多种方法划分为基于几何的技术、面 向像素的技术、基于图标的技术、基于层次的技术、基于图像的技术和分布式技术等方法。

**2.** **数据可视化技术的特点**

利用数据可视化技术，通过分析大量、复杂和多维的数据，能够提供像人眼一样的直 觉的、交互的和反应灵敏的可视化环境。

数据可视化技术的特点如下所述。

(1)交互性

用户可以方便地以交互的方式管理和开发数据。



**大数据基础与应用**

(2)多维性

可以看到表示对象或事件的数据的多个属性或变量，而数据可以按其每一维的值，将 其分类、排序、组合和显示。

(3)可视性

数据可以用图像、曲线、二维图形、三维体和动画来显示，并可对其模式和相互关系 进行可视化分析。

**3.** **数据可视化的应用**

由于数据可视化所处理的数据量十分庞大，生成图像的算法又比较复杂，过去常常需 要使用巨型计算机和高档图形工作站等。因此，数据可视化开始都在国家级研究中心、高 水平的大学、大公司的研发中心进行研究和应用。近年来，随着 PC 功能的提高、各种图 形显卡以及可视化软件的发展，可视化技术已扩展到科学研究、军事、医学、经济、工程 技术、金融、通信和商业等各个领域。下面较简单地介绍几种典型的应用领域。

(1)医学

医学数据的可视化是数据可视化领域中最为活跃的应用领域之一。由于近代非侵入诊 断技术如CT 、MRI 和正电子放射断层扫描 (PET) 的产生与发展，医生已经可以较易获得 病人有关部位的一组二维断层图像。CT 打破传统的胶片感光成像模式，通过计算机重构人 体器官或组织的图像，使医学图像从二维走向三维，使人们从人体外部可以看到内部。PET 把核技术与计算机技术结合起来。经核素标记的示踪剂注入人体后，核素衰变过程中产生 的正电子湮灭通过电子检测和计算机重构成像，使我们可以得到人体代谢或功能图像。在 此基础上，利用可视化软件，对上述多种模态的图像进行图像融合，可以准确地确定病变 体的空间位置、大小、几何形状以及它与周围生物组织之间的空间关系，从而及时、高效 地诊断疾病。

由于 EBCT 血管造影图像时间分辨率高，消除了呼吸及运动伪影，可以明确诊断各种 主动脉病变和显示冠状动脉搭桥血管解剖结构。三维重建图像利于整体直观地显示病变， 帮助明确诊断并指导手术，从而在主动脉病变的诊断和冠状动脉搭桥术后的血管显示方面， 可望取代有创的常规血管造影。

在可视化技术的基础上可以进一步实现放射治疗、矫形手术等的计算机模拟及手术规 划。例如，在做脑部肿瘤放射治疗时，需要在颅骨上穿孔，然后将放射性同位素准确地安 放在脑中病灶部位，既要使治疗效果最好，又要保证整个手术过程及同位素射线不伤及正 常组织。由于人脑内部结构十分复杂，而且在不开颅的情况下，医生无法观察到手术实际 进行情况，因而要达到上述要求是十分困难的。利用可视化技术就可以在重构出的人脑内 部结构三维图像的基础上，对颅骨穿孔位置、同位素置入通道、安放位置及等剂量线等进 行计算机模拟，并选择最佳方案。同时还可以在屏幕上监视手术进行的情况，从而大大提 高手术的成功率。又如，有不少儿童的髋关节发育不正常，当做矫形手术时，需要对髋关 节进行切割、移位、固定等操作。利用可视化技术可以首先在计算机上构造出髋关节的三 维图像，然后在计算机上对切割部位、切割形状、移位多少及固定方式等的多种方案进行 模拟，从而大大提高矫形手术的质量。





**第13章** **大数据可视化**

(2)油气勘探

在油气勘探中，除了寻找新油田之外，还需要通过改善分析和回收方法，使现存油田 处于最佳状态，并延长油田的产油寿命。科学家和工程技术人员必须先对大量的地震勘探 数据进行精确的解释，然后才能确定有无石油，并确定对地下资源的开采管理方案。油气 勘探的主要方式是通过天然地震波或人工爆炸产生的声波在地质构造中的传播，来重构大 范围内的地质构造，并通过测井数据了解局部区域的地层结构，探明油藏气藏位置及其分 布，估计蕴藏量及其勘探价值。由于地震数据及测井数据的数据量极其庞大，而且分布不 均匀，因而无法根据纸面上的数据做出分析。利用可视化技术可以从大量的地质勘探数据 或测井数据中，构造出感兴趣的等值面、等值线，并显示其范围及走向。利用不同颜色显 示出多种参数及其相互关系，从而使专业人员能对原始数据做出正确解释，得到油藏是否 存在、油藏位置及储量大小等重要信息。这不仅可以指导打井作业、减少无效井位、节约 资金，而且大大提高寻找油藏的效率，从而具有重大的经济效益及社会效益。数据的可视 化软件已在全世界许多油田和天然气开发中得到广泛的应用。利用这种软件，可以进行地 震数据处理、测井多井评估、模拟油气的储存和生产过程。不仅能确定油气储存的位置， 而且可以跟踪油气的运动，便于确定开采油气的最优路径。

(3)气象预报

气象预报关系到亿万人民的生活、国民经济的持续发展和国家安全。对灾害性天气的 预报和预防将大大减少人民生命财产的损失。气象预报的准确性依赖于对大量数据的计算 和对计算结果的分析。 一方面，科学计算可视化可将大量的数据转换为图像，在屏幕上显 示出某一时刻的等压面、等温面、旋涡、云层的位置及运动、暴雨区的位置及其强度、风 力的大小及方向等，使预报人员能对未来的天气做出准确的分析和预测。另一方面，根据 全球的气象监测数据和计算结果，可将不同时期全球的气温分布、气压分布、雨量分布及 风力风向等以图像形式表示出来，从而对全球的气象情况及其变化趋势进行研究和预测。 气象预报软件系统的关键部分是显示天气数据的三维图像，利用这个系统可以将从气球、 地面站、雷达、飞机和卫星等收集来的大量数据进行显示和处理，并在此基础上及时跟踪 和评估当地的重要气象情况，从而及时准确地做出天气预报。通常情况下，气象工作者将 二维的层状数据人为叠加来进行分析，而运用三维可视化，可让气象工作者从大量二维图 像计算中解脱出来，让他们的精力集中于预报所需的实际数值。

(4)工程

计算机辅助工程包括计算机辅助设计 (CAD) 、 计算机辅助制造 (CAM) 和计算机辅 助运行等多项内容。可视化技术有助于整个工程过程一体化，并能使技术人员看到与了解 整个过程中参数变化对整体的动态影响，从而达到缩短研制周期、节省工程费用的目的。 可视化技术可将多种来源的各种数据，包括表格数据、离散采样数据、贴体坐标数据、多 重半结构数据和非结构数据等，融合成三维的图形图像。

计算力学在工程设计中广泛应用，但计算力学离不开可视化技术。有限元分析是20世 纪50年代提出的适用于计算机处理的一种结构分析的数值计算方法。有限元分析在飞机设 计、水坝建造、机械产品设计、建筑结构应力分析等领域都得到了广泛应用。从数学的观 点来看，有限元分析将研究对象划分为若干个子单元，并在此基础上求出偏微分方程的近



**大数据基础与应用**

似解。在有限元分析中，应用可视化技术可以实现形体的网格划分及有限元分析结果数据 的图形显示，即有限元分析的前后处理，并根据分析结果，实现网格划分的优化，使计算 结果更加可靠和精确。

飞机、汽车、船舶等设计都必须考虑在气体、液体高速运动的环境中获得优良性能和 正常工作问题。传统的做法是将所设计的飞机模型放在大型风洞或水洞里做流体动力学的 物理模拟实验，然后根据实验结果修改设计。这种做法不仅耗费资金，而且又延长了设计 周期。目前已实现了在计算机上进行流体动力学的模拟计算，这就是计算流体动力学。其 核心是求解表示流体流动的偏微分方程。利用超级计算机可以对复杂几何模型的 Navier-Stokes 方程式求解。最后可计算出流场中各种参数在各个时刻的数值，但数据量十 分庞大。为了理解和分析流体流动的模拟计算结果，必须利用可视化技术在屏幕上将数据 动态地显示出来。例如，用多种不同方法表示出每一点的速度、压力、温度和组分等，并 显示出涡流、冲击波、剪切层、尾流及湍流等。在流场的可视化中，既要提高显示速度， 又要逼真地显示流场的细微结构和各种参数的等值面。当然，计算流体动力学和有限元分 析一样，计算的速度和准确度受网格划分的影响很大，通过可视化技术可以针对不同对象， 找到最适合的网格划分方法。在航空航天领域的数字模拟设备，不仅将可视化技术用于CFD 计算，同时也用于从风洞试验获得的二维图像重构三维流场，并进行计算结果与试验结果 的比较分析。特别是利用基于高度三维交互特性的虚拟现实技术，构筑虚拟风洞，为分析 各种非定常流动中的复杂结构，提供直观的研究环境。

**13.2** **大数据科学可视化**

传统的科学可视化技术趋于成熟，并已经成功应用于各学科领域之中。科学可视化技 术可以通过验证有效性和实用性来得以完善。但是，如果将这些传统的科学可视化方法直 接应用于大数据科学可视化中，将面临一系列的技术问题。这是由于大数据的特征所决定 的，而且科学可视化方法往往只有在应用数据规模达到一定程度的时候才会出现问题，从 而我们必须对现存的科学可视化技术的重新审视与深入研究，而不能直接应用于大数据科 学可视化需求。

**13.2.1** **高可伸缩性的分布式并行可视化算法**

可伸缩性又称为可扩展性，是构造分布式并行算法的重要指标。早期的可视化算法设 计基本上是面向单台计算机或单处理单元，但是单计算机或单处理单元无法实时处理 TB 数量级和更大数量级的数据。实时处理大数据需要多台计算机组成的计算机集群。

传统的科学可视化算法都应用在小规模的计算机集群中， 一般包括数十或者上百个计 算节点。但是随着计算机科学与技术的发展，实际应用有可能要求科学可视化算法能运行 在数千甚至上万个节点上，这就对科学可视化算法的可扩展性提出了更高的要求，需要设计 分布式并行可视化算法。其中并行效率及其可伸缩性是分布式并行可视化算法的主要指标。





**第13章** **大数据可视化**

某些算法是数据分析流程中的关键部分，但是本身很难并行化。在数据规模较小的情 况下，对这些算法的伸缩性要求并不高。但是随着数据规模的增大，这些算法的效率逐渐 成为数据分析流程的瓶颈，需要设计新的分布式并行算法。

**13.2.2** **并行图像合成算法**

传统的并行绘制算法，即并行图像合成算法，可以分为下述三类。

● 先分割 (sort-first) 算法

● 中间分割 (sort-middle ) 算法

● 后分割 (sort-last) 算法

其中，由于后分割算法可扩展性好，已被广泛应用于大规模并行绘制中，该算法主要 步骤如下。

① 将数据分割并分配到每台计算节点上。

② 每台计算节点独立绘制各自所分配到的数据，在这一步节点之间基本不需要数据 传输。

③ 所有计算节点将各自绘制的图片汇总并合成最终的图片结果。

其中，第三步是并行图像合成。由于节点间可能需要大量的数据交换，所以这一步往 往成为并行绘制的瓶颈。解决这类问题的关键在于减少计算节点之间的通信开销，其基本 思路是对数据进行划分并在各计算节点间进行分配。而采用的划分和分配方案，需要与数 据的访问一致，即如何让节点只使用自身的数据进行颗粒跟踪，进而减少数据交换。

**13.2.3** **并行颗粒跟踪算法**

传统的大规模科学可视化研究对象主要集中在三维标量场数据。但是，在实际中，很 多科学模拟使用和输出三维流场数据。虽然流场可视化是一个很活跃的研究领域，但对于 大规模三维流场可视化的研究却相对较少。

**1.** **大规模三维流场可视化的研究原因**

首先，因为很容易造成视觉上的杂乱无序，所以大多数二维的流场可视化方法不能直 接应用在三维流的结构。然而，颗粒跟踪是一个非常难以并行处理的算法。每个颗粒虽然 可以单独跟踪，但是可能出现在空间中的任何一个位置，这就需要计算节点之间通过通信 交换颗粒。同时，当大量的颗粒在空间移动时，每个计算节点可能会处理不同数量的颗粒， 从而造成计算量严重失衡。解决这些问题的关键在于如何减少计算节点之间的通信开销， 其基本思路是对数据进行划分和在计算节点间进行分配，而采用的划分和分配方案，需要 与数据的访问一致，即如何让节点只使用自身的数据进行颗粒跟踪，而减少数据交换。基 于此思路产生了层次可视化方法以实现并行颗粒跟踪。

**2.** **层次可视化方法的过程**

层次可视化方法分为两步：第一步对流场数据进行预处理，将数据分割、分配到计算



**大数据基础与应用**

节点；第二步，进行并行颗粒跟踪。在第一步预处理中，可将三维时变流场用四维向量场 表示，然后进行采样，生成采样后的四维向量场。对于这个小规模的向量场，使用向量场 聚类分析方法，生成层次聚类结构。每一层的聚类结果，代表了对向量场的分割，而且每 个分割区域与向量场的方向一致。也就是说，如果在每个区域中进行颗粒跟踪，则颗粒的 运动轨迹与区域基本保持一致。这样就很自然地提供了一种数据分割和分配方案，而且这 种方案与数据访问一致。因此，在第二步中，就可以很方便地进行颗粒跟踪，而减少数据 通信。实验显示，该算法可以提高相对于节点数的线性性能。另外，该算法可以应用在三 维静态流场，生成流线，或者应用在三维时变流场，生成轨迹线。同时，在生成的可视化 中，能够通过控制线的数目来减少视觉复杂度，同时保留可视化的正确性。传统的可视化 方法无法对该数据进行实时绘制，而并行算法具有良好的可扩展性，能够使用多台机器， 实现实时交互可视化。该算法还可以通过选择不同的层次，来控制显示的细节程度。

**13.2.4** **重要信息的提取与显示**

大数据科学可视化的另一个重要研究方向是如何从数据中快速有效地提取重要信息， 并且用这些重要信息来指导可视化的生成。从可视化的角度， 一方面需要设计可视化技术， 表达数据中特定信息的定义，通过人机交互工具，由用户来调整参数，观察和挖掘数据中 的重要信息；另一方面，根据用户的反馈信息，调整可视化，以更好地显示重要信息，淡 化非重要信息，方便用户对重要信息及其背景的观测。整个信息的提取过程是个典型的交 互式可视化分析过程。下面介绍基于这一思想的两个技术：用于流场可视化的层次流线束 技术和用于标量数据的基于距离场的可视化技术。

**1.** **层次流线束的可视化**

传统的基于流线的可视化方法常常无法避免视觉复杂度的问题。当在流场中放入过多 的流线的时候，视觉复杂度会大大提高，造成视觉混乱，如果放入过少的流线，又无法保 证流线能够覆盖一些重要的流场区域，比如旋涡中心，从而可能会丢失一部分重要的流场 结构信息。为了解决这个问题，提出层次流线束技术。

这个技术的基本思想是：

首先，让流线能够经过流场的重要区域；其次，对于选择的流线，提取其附近的流线， 使它们组成流线束，这样就可以大大增强视觉对线的辨别；再次，构造层次结构，能够控 制流线束的数目和密度，显示不同层次的细节；最后，流线束也提供了流场的分割，允许 我们拆分流场，来观测不同部分的结构。

算法的实现步骤如下。

① 通过流场的曲率场和扭矩场计算出流场中的重要部分，这些重要部分一般对应于流 场的拓扑关键结构，比如旋涡中心等。根据这些重要信息放置流线种子，生成足够多的流 线来覆盖这些区域。

② 采用传统的层次聚类分析方法，根据线与线之间的相似程度来构造层次聚类，在同 一聚类中，线条具有相同的形状。



**第13章** **大数据可视化**

③ 对每个聚类提取代表性的流线。对于每个聚类只保留在聚类边缘处的若干条流线， 这些流线具有类似的形状，从而形成线束。可以通过选择层次和调整参数，显示不同数量 的线束，或者对它们进行分割，进而对这些复杂的大规模流场进行方便的观测。

该技术可以让用户方便地挑选并集中观察少数聚类，并将它们与其他聚类区分开。 对 于选择的聚类，用户还可以控制线的密度，即线束的粗细。

**2.** **基于距离场的可视化**

基于距离场的可视化应用在许多科学领域中，常常对数据中某些区域感兴趣，例如体 数据中的某个等值面。对于这种原因，可以在体数据中的每个点上定义一个变量来表示这 个点的重要性。这个重要性的值等于这个点到感兴趣的表面距离的倒数。距离越近，越重 要；反之，不重要。然后可以根据这个重要性值，来定义体数据中每个点的透明度值。越 不重要的区域越透明，降低可见度。

通过控制不同的距离值，来控制变量在这个等值面周围的显示量。这种基于距离场的 可视化，使我们第一次看到湍流燃烧中的小的旋涡结构，而这个结构因为太靠近等值面，

过去很难用传统的可视化技术显示。上述这个思路使可以根据距离值来决定数据精度。如 果距离重要的区域近，那么就采用高精度的数据；反之，则可以使用一些低精度的数据。

**13.2.5** **原位可视化**

传统的可视化工作以后处理的模式出现，也就是在科学模拟计算结束后运行。但是， 随着计算速度的提高，I/O 速度与计算速度之间的差距越来越大，在模拟计算过程中，有的 数据在生成后来不及保存到存储器中。另一方面，计算规模越来越大，而相应生成的数据 规模也越来越大，现有的存储系统无法把所有的计算数据都保存下来。为了解决这两方面 的问题，常用方法是采用空间或者时间上的采样方法，最后只保存部分数据。但是，这个 方法造成结果数据的丢失，不能保证高精度数值模拟，无法充分利用高效能计算机的优势， 造成资源浪费。

原位可视化是解决上述问题的一个有效方法，其基本思想是将可视化与科学模拟集成 在一起。

在科学模拟的过程中，每个时间片的结果生成之后，就可以立刻调用可视化模块。为 了减少数据的冗余，可视化程序与科学模拟程序可以共享数据结构。

由于数据的分割和分配优先满足科学模拟的需求，可视化程序的工作分配有可能出现 不均衡，所以需要重现可视化的工作量在各个计算节点上分配算法，减少数据传输。

可视化程序的开销不能太高，要保持集成系统的高效能，必须提高可视化程序的效率， 其可扩展性必须与科学模拟一致，能够使用上万个、10多万个或更多的计算节点。

大规模科学数据可视化除了从科学模拟中生成图片，还能使科学模拟可以操作。可以 交互式地调整科学模拟程序的参数，并得到实时的反馈。参数调整越实时，观察不同参数 间的因果关系时就越方便，从而有助于更直观地理解模型、算法与数据。





**大数据基础与应用**

**13.3** **大数据可视化分析**

可视化分析是信息可视化与科学可视化发展的产物，侧重于借助交互式用户界面进行 的推理。主要包含分析推理，视觉呈现和交互，数据表示和转换，以及支持产生、表达和 传播分析结果的技术等四部分核心内容。可视化分析技术通过交互可视界面来进行分析、

推理和决策，从海量、动态、不确定甚至相互冲突的数据中整合信息，获取对复杂情景的 更深层的理解，可供人们检验已有预测，探索未知信息，同时提供快速、可检验、易理解 的评估和更有效的交流手段。数据可视化分析技术主要应用于大数据关联分析，由于所涉 及的信息比较分散、数据结构不统一，通常以人工分析为主，加上分析过程的非结构性和 不确定性，所以不易形成固定的分析流程或模式，很难将数据调入应用系统中进行分析挖 掘。借助功能强大的可视化数据分析平台，可辅助人工操作将数据进行关联分析，并做出 完整的分析图表。图表中包含所有事件的相关信息，完整展示数据分析的过程和数据链走 向。同时，这些分析图表也可通过另存为其他格式，供相关人员调阅。

大数据可视化分析必须采用有效的数据管理方法。这也是创建一个混合环境的需要。 在实践应用中，首先是在大数据环境下进行探索，如果这些探索揭示了某些数据所需要的 价值，那么利用各种技术得到数据分析结果之后，将各种数据分析结果用形象直观的方式 展示出来，如标签图、气泡图、雷达图、热力图、树形图、辐射图、趋势图等都是可视化 的表现方式，这样用户能够快速发现数据中蕴含的规律特征。

**13.3.1** **大数据可视化分析概念**

可视化分析是一个新的学科方向，是指通过交互可视界面来进行分析、推理和决策的 科学。可视化分析与各个领域的数据形态、大小及其应用密切相关。

可视化分析领域关注的关于人类感知与用户交互的问题。由于大数据改变了人类的工 作生活方式，研究者开始寻找有关大数据问题的可视分析解决方案。大数据来自不同领域 的科学、工程、社会和网络的模拟与观察实测。尽管还有许多PB 量级甚至TB 量级规模的 数据分析问题还没有解决，但科学家现已开始研究 EB 量级的数据。大数据可视分析通常 结合了用于计算的高性能计算机群、处理数据存储与管理的高性能数据库组件及云端服务 器和提供人机交互界面的桌面计算机。

**13.3.2** **大数据可视化分析方法**

**1.** **原位交互分析技术**

在进行可视化分析时，将在内存中的数据尽可能多地进行分析称为原位交互分析。

对于超过 PB 量级以上的数据，先将数据存储于磁盘，然后读取进行分析的后处理方 式已不适合。与此相反，可视分析则在数据仍在内存中时就会做尽可能多的分析。这种方 式能极大地减少 I/O 的开销，并且可实现数据使用与磁盘读取比例的最大化。应用原位交 互分析出现下述问题：





第13章 大数据可视化

① 使得人机交互减少，进而容易造成整体工作流的中断。

② 硬件执行单元不能高效地共享处理器，导致整体工作流的中断。

**2.** **数据存储技术**

大数据是云计算的延伸，云服务及其应用的出现影响了大数据存储。流行的 Apache Hadoop 架构已经支持在公有云端存储 EB 量级数据的应用。许多互联网公司，如 Facebook、 谷歌、eBay 和雅虎等，都已经开发出了基于 Hadoop 的 EB 量级的超大规模数据应用。 一 个基于云端的解决方案可能满足不了 EB 量级数处理。 一个主要的疑虑是每 GB 的云存储 成本仍然显著高于私有集群中的硬盘存储成本。另一个问题是基于云的数据库的访问延时 和输出始终受限于云端通信网络的带宽。不是所有的云系统都支持分布式数据库的 ACID 标准。对于 Hadoop 软件的应用，这些需求必须在应用软件层实现。

**3.** **可视化分析算法**

传统的可视化分析算法设计没有考虑可扩展性，因此，传统算法的特点是计算过于复 杂，或者输入的不易理解的从分简明结果。并且，大部分算法都附设了后处理模型的假设， 认为所有数据都在内存或本地磁盘中可被直接访问。对于大数据的可视化算法不仅要考虑 数据大小，而且要考虑视觉感知的高效算法。需要引入创新的视觉表现方法和用户交互手 段。更重要的是用户的偏好和习惯必须与自动学习算法有机结合起来，这样可视化的输出 具有高度适应性。为了减少数据分析与探索的成本及降低难度，可视化算法应具有巨大的 控制参数搜索空间，进而自动算法可以组织数据并且减少搜索空间。

**4.** **数据移动、传输和网络架构**

随着计算成本的下降，数据移动成本已成为可视化分析中付出代价最高的部分。由于 数据源常常分布在不同的地理位置，并且数据规模巨大，高效实现构成了绝大多数大规模 模拟系统中代码的基石。由于可视化分析计算将运行在更大的系统上，必须提出更加有效 的算法、开发更加高效的软件，能够有效地利用网络资源，并且能提供更加方便通用的接 口，使得可视化分析的专家能高效地数据挖掘。

**5.** **不确定性的量化**

如何量化不确定性已经成为许多科学与工程领域的重要问题。了解数据中不确定性的 来源，对于决策和风险分析十分重要。随着数据规模增大，直接处理整个数据集的能力也 受到了极大的限制。许多数据分析任务中引入数据亚采样来应对实时性的要求，由此也带 来了更大的不确定性。不确定性的量化及可视化对未来的可视化分析工具而言极为重要， 必须发展可应对不完整数据的分析方法，许多现有算法必须重视设计，进而考虑数据的分 布情况。 一些新兴的可视化技术会提供一个不确定性的直观视图，来帮助用户了解风险， 从而帮助用户选择正确的参数，减少产生误导性结果的可能。从这个方面来看，不确定性 的量化与可视化将成为绝大多数可视化分析任务的核心。

**6.** **并行计算**

并行处理可以有效地减少可视计算所用的时间，从而实现数据分析的实时交互。未来





**大数据基础与应用**

的计算体系结构将在一个处理器上置入更多的核，每个核所占有的内存也将减少，在系统 内移动数据的代价也会提高。大规模并行化甚至可能出现在桌面 PC 或者笔记本电脑平台 上。并行计算的普及就在不远的将来。为了发掘并行计算的潜力，许多可视化分析算法需 要完全地重新设计。在单个核心内存容量的限制之下，不仅需要有更大规模的并行，也需 要设计新的数据模型。需要设计出既考虑数据大小又考虑视觉感知的高效的算法，需要引 入创新的视觉表现方法和用户交互手段。更重要的是，用户的偏好和习惯必须与自动学习 算法有机结合起来，这样可视化的输出才具有高度适应性。当可视化算法拥有巨大的控制 参数搜索空间时，自动算法可以组织数据并且减少搜索空间，这对于减少数据分析与探索 的成本和降低难度起着关键的作用。

**7.** **面向领域与开发的库、框架以及工具**

由于缺少低廉的资源库、开发框架和工具，基于高性能计算的可视化分析应用的快速 研发受到了严重的阻碍。这些问题在许多应用领域十分普遍，比如用户界面、数据库以及 可视化，而这些领域对于可视化分析系统的开发都是至关重要的。在绝大部分的高性能计 算平台上，即使是最基本的软件开发工具，也是罕见的。这种资源的稀缺对于科学领域的 用户来说是令人十分沮丧的。许多在桌面平台上流行的可视化和可视化分析软件，如果放 到高性能计算平台上，不是太昂贵就是有待开发，而为高性能计算平台开发这样定制的软 件，则是个耗时耗力的做法。

**8.** **用户界面与交互设计**

由于传统的可视化分析算法的设计通常没有考虑可扩展性，所以许多算法的计算过于 复杂或者不能输出易理解的简明结果。又由于数据规模不断地增长，以人为中心的用户界 面与交互设计面临多层次性和高复杂性的困难。计算机自动处理系统对于需要人参与判断 的分析过程的性能不高，现有的技术不能更充分发挥人的认知能力。利用人机交互可以化 解上述问题。为此，在大数据的可视化分析中，用户界面与交互设计成为研究热点，主要 应考虑下述问题。

(1)用户驱动的数据简化

在数据量巨大的情况下，通过压缩来简化数据的传统方法已变得无效。需要让用户根 据他们的数据收集情况与分析需求方便地控制简化过程。

(2)可扩展性与多级层次

在可视化分析中，解决可扩展性问题的主要方法是多层次办法。但是当数据量增大时， 层级的深度与复杂性也随之增大。在继承关系复杂且深度大的层次关系中，搜索最优解是 可扩展性分析的研究问题。

(3)表示证据和不确定性

一个可视化分析环境中，表示证据与不确定性量化通常得到统一，并且需要人的参与 和诠释。需要研究如何通过可视化来清晰地表示证据和不确定性。

(4)异构数据融合

大数据通常都是高度异构的。因此，在分析异构数据中的对象或实体的相互关系上需



**第13章** **大数据可视化**

要花费很大功夫。面临的问题是如何从大数据中抽取出合适数量的语义信息，将其交互地 融合后进行可视化分析。

(5)交互查询中的数据概要与分流

当数据规模超过了PB 量级时，对整个数据集进行分析通常不现实，也是没有必要的。 数据的概要与分流使得用户能够请求满足特定特性的数据子集，而它面临的挑战是让I/O 部 件能在数据概要与分流的结果中顺利运行，从而使得用户能对超大规模数据进行交互查询。

(6)时变特征分析

一个超大规模的时变数据集通常在时间上延续很长，而在频谱上或者空间上，依据数 据集类型较窄。主要的问题是要开发有效的可视化分析技术，不仅在计算上是可行的，同 时也能最大限度地发掘在追踪数据动态变化特征上的人的认知能力。

(7)设计与工程开发

对于系统开发者来说，他们缺少在高性能计算平台上的社区尺度应用程序接口和框架 支持。高性能计算社区必须为高性能计算系统上的用户界面与交互的开发建立规范的设计 和提供工程资源。

拥有大量的数据并不等于获得相应数据的价值。只有能够有效理解数据，才能真正利 用好大数据。复杂数据，例如微博的用户关系、庞杂的事件发展对发展相应的分析工具提 出了新的挑战和要求。可视化和可视化分析与其他分析手段不同，可视化利用了人类视觉 认知的高通量特点，通过图形的形式表现信息的内在规律及其传递、表达的过程，是人们 理解复杂现象，诠释复杂数据的重要手段和途径。可视化和可视化分析技术也越来越广泛 地被应用到科学、工程、商业和日常生活中。利用可视化与可视化分析技术，通过交互可 视界面的分析、推理和决策，从海量、动态、不确定甚至相互冲突的数据中整合信息，获 取对复杂情景的更深层的理解，可供人们检验已有预测，探索未知信息，同时提供快速、 可检验、易理解的评估和更有效的交流手段。可视化能够帮助科学家理解超新星的爆发、 地震发生的规律、大都市交通的拥堵、发掘微博扩散传播的路线以及构建对复杂信息的可 视化分析平台。可视化科学技术应用广泛，前程远大。

小结

数据可视化技术利用图形、图像处理、计算机视觉以及用户界面，通过表达、建模以 及对立体、表面、属性以及动画的显示，对数据加以可视化解释。科学可视化的主要过程 是建模和渲染。

信息可视化是指跨学科领域的大规模非数值型信息资源的视觉展现，帮助人们理解和 分析数据。数据可视化技术是指运用计算机图形学和图像处理技术，将数据转换为图形或 图像在屏幕上显示出来，并利用数据分析和开发工具发现其中未知信息的交互处理的理论、 方法和技术。

数据可视化概念已大大扩展，不仅包括科学计算数据的可视化，而且包括工程数据和 测量数据的可视化。通常又把这种空间数据的可视化称为体视化技术。数据可视化技术具 有交互性、多维性和可视性等特点。可视化分析科学是指通过交互可视界面来进行分析、 推理和决策的科学。



云计算中用户的安全需求

威胁模型

云安全的支撑技术

云安全

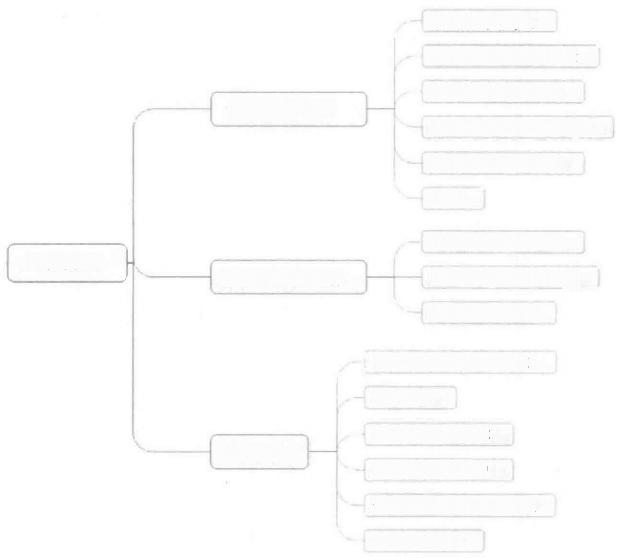
用户数据隐私保护

云计算执行环境的可信性

资源共享问题

**第14章** **大数据安全**

本章主要内容



数据安全的定义

数据处理与存储的安全

数据安全的基本特点

大数据安全概述

威胁数据安全的主要因素

安全制度与防护技术

应用

大数据的不安全因素

大数据安全

大数据安全的关键问题

大数据安全措施

大数据安全的内容

计算机系统安全是指为数据处理系统建立和采用的技术和管理的安全保护，保护计算 机硬件、软件和数据不因偶然和恶意的原因遭到破坏、更改和泄露。计算机网络的安全是 指通过采用各种技术和管理措施，使网络系统正常运行，从而确保网络数据的可用性、完 整性和保密性。所以，建立网络安全与计算机系统安全保护措施的目的是确保经过网络传 输和交换的数据不发生增加、修改、丢失和泄露等，实质上是保证数据的安全。

**14.1** **大数据安全概述**

数据具有普遍性、共享性、增值性、可处理性和多效用性等特性，数据资源具有特别 重要的意义与价值。数据安全就是要保护信息系统或网络中的数据资源免受各种类型的威 胁、干扰和破坏，所以数据安全的研究意义非凡。



第14章 大数据安全

**14.1.1** **数据安全的定义**

数据安全包括数据本身的安全和数据防护的安全两个方面的内容。

**1.** **数据本身的安全**

数据本身的安全是指采用密码算法对数据进行主动保护，如数据保密、数据完整性、 双向强身份认证等。

**2.** **数据防护的安全**

数据防护的安全主要是采用现代信息存储手段对数据进行主动防护，如通过磁盘阵列、 数据备份、异地容灾等手段保证数据的安全。

数据安全是一种主动的措施，数据本身的安全必须基于可靠的加密算法与安全体系， 主要是有对称算法与公开密钥密码体系两种。

**14.1.2** **数据处理与存储的安全**

**1.** **数据处理的安全**

数据处理的安全是指如何有效地防止数据在录入、处理、统计或打印中由于硬件故障、 断电、死机、人为的误操作、程序缺陷、病毒或黑客等造成的数据库损坏或数据丢失现象。 某些敏感或保密的数据可能被不具备资格的人员或操作员阅读，进而造成数据泄密等后果。

**2.** **数据的存储安全**

数据存储的安全是指数据库在系统运行之外的可读性，对于一个标准的Access 数据库， 很容易打开阅读或修改。 一旦数据库被盗，即使没有原来的系统程序，也可以另外编写程 序对盗取的数据库进行查看或修改。从这个角度说，不加密的数据库是不安全的，容易造 成商业泄密。这就需要考虑计算机网络通信的保密、安全及软件保护等问题。

**14.1.3** **数据安全的基本特点**

数据安全具有保密性、完整性和可用性三个基本特点。

**1.** **保密性**

保密性又称为机密性，是指个人或团体的信息不为其他不应获得者而获得。在计算机 中，许多软件包括邮件软件、网络浏览器等，都有保密性相关的设定，用以维护用户信息 的保密性，另外间谍或黑客也有可能造成保密性的问题。

**2.** **完整性**

数据完整性是指在传输、存储数据的过程中，确保数据不被没有授权者篡改，或在篡



**大数据基础与应用**

改后能够被迅速发现。在信息安全领域中，常常与保密性边界混淆。以普通 RSA 对数值信 息加密为例，黑客或恶意用户在没有获得密钥破解密文的情况下，可以通过对密文进行线 性运算来改变数值信息的值。例如交易金额为A 元，通过对密文乘3,可以使交易金额成 为 3A。为了解决上述问题，通常可以使用数字签名或散列函数对密文进行保护。

**3.** **可用性**

数据可用性是以使用者为中心的设计概念，易用性设计在于使产品的设计能够符合使 用者的习惯与需求，达到易用。例如在网站设计中，主要考虑的问题之一是使用者在浏览 的过程中不产生压力或感到挫折，并能使使用者在使用网站功能时，能用最少的努力，发 挥最大的效能。既然可用性是数据安全的三个特点之一，那么有违数据的可用性就是违反 数据安全的规定。

**14.1.4 威胁数据安全的主要因素**

威胁数据安全的主要因素如下所述。

**1.** **数据信息存储介质的损坏**

在物理介质层次上对存储和传输的信息进行安全保护，是信息安全的基本保障。物理 安全隐患大致包括下述三个方面。

① 自然灾害(如地震、火灾、洪水、雷电等)、物理损坏(如硬盘损坏、设备使用到 期、外力损坏等)和设备故障(如停电断电、电磁干扰等);

② 电磁辐射、信息泄露、痕迹泄露(如口令密钥等保管不善);

③ 操作失误(如删除文件、格式化硬盘、线路拆除)、意外疏漏等。

**2.** **人为因素**

人为因素包括人为的无意失误和人为恶意攻击。

网络管理员安全配置不当造成的安全漏洞，用户安全意识不强，口令选择不慎，用户 将自己的账号随意转借他人或与别人共享等都将对网络信息安全带来威胁。别有用心的人 将利用这些无意的失误，从他人那里获取不该他获取的信息。

恶意攻击是计算机网络所面临的最大威胁，网络战中敌方的攻击和计算机犯罪就属于 这一类。这类攻击又可分为两种： 一种是主动攻击，是以各种方式有选择地破坏信息的有 效性和完整性；另一种方式是被动攻击，就是在不影响网络正常工作的情况下，进行截获、 窃取、破译以获得重要机密信息。这两种攻击均可对计算机网络造成极大的危害，并造成 机密数据的泄露。

由于操作失误，使用者可能将误删除系统的重要文件，或者修改影响系统运行的参数， 以及没有按照规定要求或操作不当导致的系统停机。

随着计算机网络信息系统日益复杂，以致人们无法保证系统不存在网络设计漏洞与管 理漏洞。这些漏洞和缺陷自然成为黑客进行攻击的首选目标。另外，软件设计人员出于自 身的考虑，在进行软件开发时，为所开发的软件设置“后门”,一旦“后门”为外人所知，





第14章 大数据安全

该软件则无安全可言，所造成的后果也不堪设想。

**3.** **黑客**

电脑入侵、账号泄露、资料丢失、网页被黑等也是企业信息安全管理中经常遇到的问 题。其特点是往往具有明确的目标。当黑客要攻击一个目标时，通常是首先收集被攻击方 的有关信息，分析被攻击方可能存在的漏洞，然后建立模拟环境，进行模拟攻击，测试对 方可能的反应，再利用适当的工具进行扫描，最后通过已知的漏洞，实施攻击。然后就可 以读取邮件，搜索和盗窃文件，毁坏重要数据，破坏整个系统的信息，造成不堪设想的后 果。入侵者通过网络远程入侵系统，主要的侵入形式：系统漏洞、管理不力等。

**4.** **病毒**

计算机病毒能影响计算机软件、硬件的正常运行，破坏数据的正确与完整，甚至导致 系统崩溃等重大恶果，特别是一些针对盗取各类数据信息的木马病毒等。目前杀毒软件普 及较广，计算机病毒造成的数据信息安全威胁隐患已经缓解很多。由于感染计算机病毒而 破坏计算机系统，造成的重大经济损失屡屡发生，计算机病毒的复制能力强，感染性强， 特别是网络环境下，传播性更快。

**5.** **信息窃取**

复制、删除计算机上的信息或偷走计算机。

**6.** **电源故障**

电源供给系统故障， 一个瞬间过载电功率将损坏在硬盘或存储设备上的数据。

**7.** **磁干扰**

磁干扰是指数据接触到有磁性的物质后，将造成数据被破坏。

**14.1.5** **安全制度与防护技术**

**1.** **安全制度**

不同的单位和组织，都有自己的网络信息中心，为确保信息中心、网络中心机房重要 数据的安全， 一般要根据国家法律和有关规定制定适合本单位的数据安全制度。

① 对应用系统使用、产生的介质或数据按其重要性进行分类，对存放有重要数据的介 质，应备份必要份数，并分别存放在不同的安全地方(防火、防高温、防震、防磁、防静 电及防盗),建立严格的保密保管制度。

② 保留在机房内的重要数据，应为系统有效运行所必需的最少数量，除此之外不应保 留在机房内。

③ 根据数据的保密规定和用途，确定使用人员的存取权限、存取方式和审批手续。 ④ 重要数据库，应设专人负责登记保管，没有经批准，不得随意挪用重要数据。

⑤ 在使用重要数据期间，应严格按国家保密规定控制转借或复制，需要使用或复制的





**大数据基础与应用**

须经批准。

⑥ 对所有重要数据应定期检查，要考虑介质的安全保存期限，及时更新复制。损坏、 废弃或过时的重要数据应由专人负责消磁处理，秘密级以上的重要数据在过保密期或废弃 不用时，要及时销毁。

⑦ 机密数据处理作业结束时，应及时清除存储器、联机磁带、磁盘及其他介质上有关 作业的程序和数据。

⑧ 机密级及以上秘密信息存储设备不得并入互联网。重要数据不得外泄，重要数据的 输入及修改应由专人来完成。重要数据的打印输出及外存介质应存放在安全的地方，打印 出的废纸应及时销毁。

**2.** **防护技术**

计算机存储的数据越来越多，而且越来越重要，为了防止计算机中的数据意外丢失， 一般都采用安全防护技术来确保数据的安全，下面简单介绍常用的数据安全防护技术。

(1)磁盘阵列

磁盘阵列是指把多个类型、容量、接口甚至品牌一致的专用磁盘或普通硬盘连成一个 阵列，使其以更快的速度、准确、安全的方式读写磁盘数据，从而保证数据读取速度和安 全性。

(2)数据备份

备份管理包括备份的可计划性、自动化操作、历史记录的保存或日志记录。 (3)双机容错

双机容错的目的在于保证系统数据和服务的在线性。当某一系统发生故障时，仍然能 够正常地向网络系统提供数据和服务，使得系统不至于停顿，双机容错的目的在于保证数 据不丢失和系统不停机。

(4)NAS

NAS 解决方案通常配置为文件服务的设备，由工作站或服务器通过网络协议和应用程 序来进行文件访问，大多数 NAS 链接在工作站客户机和 NAS 文件共享设备之间进行。这 些链接依赖企业的网络基础设施来正常运行。

(5)数据迁移

由在线存储设备和离线存储设备共同构成一个协调工作的存储系统，该系统在在线存 储和离线存储设备间动态地管理数据，使得访问频率高的数据存放于性能较高的在线存储 设备中，而访问频率低的数据存放于较为廉价的离线存储设备中。

(6)异地容灾

异地实时备份是高效、可靠的远程数据存储。在IT 系统中，必然有核心部分，通常称 之为生产中心，往往给生产中心配备一个备份中心，该备份中心是远程的，并且在生产中 心的内部已经实施了各种各样的数据保护。不管怎么保护，当火灾、地震这种灾难发生时， 一旦生产中心瘫痪了，备份中心将接管生产，继续提供服务。

(7)SAN

SAN 允许服务器在共享存储装置的同时仍能高速传送数据。这一方案具有带宽高、可







**第14章** **大数据安全**

用性高、容错能力强的优点，而且它可以轻松升级，容易管理，有助于改善整个系统的总 体成本状况。

(8)数据库加密

对数据库中数据加密是为增强普通关系数据库管理系统的安全性，提供一个安全适用 的数据库加密平台，对数据库存储的内容实施有效保护。通过数据库存储加密等安全方法 实现了数据库数据存储保密和完整性要求，使得数据库以密文方式存储并在密态方式下工 作，确保了数据安全。

(9)硬盘安全加密

硬盘维修商根本无法查看经过安全加密的故障硬盘，绝对保证了内部数据的安全性。 硬盘发生故障更换新硬盘时，全自动智能恢复受损坏的数据，有效防止企业内部数据因硬 盘损坏、操作错误而造成的数据丢失。安全技术包含下述3类。

① 隐藏；

② 访问控制；

③ 密码学。

**14.1.6** **应用**

**1.** **数字水印**

数字水印是指把特定的信息嵌入到数字信号中，数字信号可能是音频、图片或影片等。 如果要复制具有数字水印的信号，所嵌入的信息也将一并被复制。

数字水印技术可分为浮现式的水印和隐藏式的水印两种类型。

(1)浮现式的水印

浮现式的水印是可见的水印，浮现式的水印信息是可见的，

即在观看图片或影片时可以同时看见水印信息。 一般来说，浮现

式的水印通常包含版权拥有者的名称或者标志。电视播放画面边

缘或角落中所放置的标志，就是浮现式水印的一种。如图14-1所

示为浮现式水印的图片，上面的文字“2006”位于图片的正中央。

(2)隐藏式的水印 图14-1 浮现式的水印

隐藏式的水印是以数字数据的方式加入音频、图片或影片中，

但是，在一般的状况下无法被看见。隐藏式水印的重要应用之一是保护版权，期望能借此 避免或阻止数字媒体没有经授权的复制。隐写技术也是数字水印的一种应用，双方可利用 隐藏在数字信号中的信息进行沟通。数字照片中的注释数据能记录照片拍摄的时间、使用 的光圈和快门，甚至是相机的厂牌等信息，这也是数字水印的应用之一。

**2.** **数字签名**

数字签名又称为公钥数字签名、电子签章。数字签名属于密码学，要避免下述的错误 理解，数字签名不是把签名扫描成数字图像，或者用触摸板获取的签名，更不是个人的落 款。数字签名的文件的完整性很容易验证，不需要骑缝章、骑缝签名，而且数字签名具有





**大数据基础与应用**

不可抵赖性，不需要笔迹专家来验证。

(1)主要功能

数字签名的功能类似写在纸上的普通的物理签名，使用了公钥加密领域的技术实现鉴 别数字信息的方法。 一套数字签名通常定义两种互补的运算， 一个运算用于签名，另一个 运算用于验证。保证信息传输的完整性、发送者的身份认证、防止交易中的抵赖发生。

(2)数字签名过程

数字签名是个加密的过程，数字签名验证是个解密的过程。当发送报文时，发送方用 一个哈希函数从报文文本中生成报文摘要，然后用自己的私人密钥对这个摘要进行加密， 这个加密后的摘要将作为报文的数字签名和报文一起发送给接收方。接收方首先用与发送 方一样的哈希函数从接收到的原始报文中计算出报文摘要，接着再用发送方的公用密钥来 对报文附加的数字签名进行解密。如果这两个摘要相同，那么接收方就能确认该数字签名 是发送方的，因此数字签名能够验证信息的完整性。

(3)数字签名的功效

数字签名有两种功效： 一是能确定消息确实是由发送方签名并发出来的，因为别人假 冒不了发送方的签名。二是数字签名能确定消息的完整性。因为数字签名的特点是它代表 了文件的特征，文件如果发生改变，数字签名的值也将发生变化。不同的文件将得到不同 的数字签名。 一次数字签名涉及一个哈希函数、发送者的公钥、发送者的私钥。

**3.** **加密技术**

在各类系统中保存的企业关键数据量越来越大，保存时限也越来越长，甚至是永久性 保存。于是关键业务数据的数据安全性问题越来越突出。如何增强企业软件系统的安全性、 保密性、真实性、完整性，已成为软件开发的关注点。从保护数据的角度来看，数据安全 可以分为数据加密、数据传输安全和身份认证管理三部分。

(1)数据加密

数据加密就是根据密码算法把明文数据变换成难以识别的密文数据，通过使用不同的 密钥，可用同一加密算法把同一明文变成不同的密文，将这一过程称为加密。当需要时， 可使用密钥把密文数据还原成明文数据，将这一过程称为解密。数据加密可以实现数据的 保密性。数据加密被公认为是保护数据传输安全唯一实用的方法和保护存储数据安全的有 效方法。

(2)数据传输安全

数据传输安全是指数据在传输过程中必须要确保数据的安全性、完整性和不可篡改性。 (3)身份认证管理

身份认证的目的是确定系统和网络的访问者是否为合法用户。主要采用登录密码、代 表用户身份的物品，如智能卡、IC 卡等，或反映用户生理特征的标识鉴别访问者的身份。

在数据安全中，数据加密技术是最基本的安全技术，是信息安全的核心，最初主要用 于保证数据在存储和传输过程中的保密性。它通过变换和置换等各种方法将被保护信息置 换成密文，然后再进行信息的存储或传输，即使加密信息在存储或者传输过程中为非授权 人员所获得，也可以保证这些信息不为其所知，进而达到保护信息的目的。方法的保密性





第14章 大数据安全

直接取决于所采用的密码算法和密钥长度。

**4.** **加密算法**

在数据加密中，根据密钥类型不同，可以将加密算法分为两类：对称加密算法(属于 私钥密码体系)和非对称加密算法(属于公钥密码体系)。在对称加密算法中，数据加密和 解密采用的都是同一个密钥，因而其安全性由所持有密钥的安全性所决定。对称加密算法 的主要优点是加密和解密速度快，加密强度高，且算法公开，但其最大的缺点是实现密钥 的秘密分发困难，在大量用户的情况下密钥管理复杂，而且无法完成身份认证等功能，不 便于应用在网络开放的环境中。最著名的对称加密算法有数据加密标准DES 和欧洲数据加 密标准 IDEA 等，加密强度最高的对称加密算法是高级加密标准 AES。对称加密算法和非 对称加密算法可以分别应用于数据加密、身份认证和数据安全传输。

(1)对称加密算法

在对称加密算法中，数据发信方把明文(原始数据)和加密密钥一起经过特殊加密算 法加密后，使其变成复杂的密文发送出去。收信方收到密文之后，则需要使用加密用过的 密钥及相同算法的逆算法对密文进行解密，才能使其恢复成可读明文。在对称加密算法中， 使用的密钥只有一个，发收信双方都使用这个密钥对数据进行加密和解密，这就要求解密 方事先必须知道加密密钥。对称加密算法的特点是算法公开、计算量小、加密速度快、加 密效率高。不足之处是，交易双方都使用同样钥匙，安全性得不到保证。此外，每对用户 每次使用对称加密算法时，都需要使用其他人不知道的唯一钥匙，这将使得钥匙数量成几 何级数增长，密钥管理成为用户的负担。对称加密算法在分布式系统上使用较为困难，主 要是因为密钥管理困难，使用成本较高。

(2)不对称加密算法

不对称加密算法使用两把完全不同但又是完全匹配的一对钥匙，即公钥和私钥。在使 用不对称加密算法时，只有使用匹配的一对公钥和私钥，才能完成对明文的加密和解密过 程。加密明文时采用公钥加密，解密密文时使用私钥才能完成，而且发信方(加密者)知 道收信方的公钥，只有收信方(解密者)才是唯一知道私钥的人。不对称加密算法的基本 原理是，如果发信方想发送只有收信方才能解读的加密信息，发信方必须首先知道收信方 的公钥，然后利用收信方的公钥来加密原文。收信方收到加密密文后，使用自己的私钥才 能解密密文。显然，采用不对称加密算法，收发信双方在通信之前，收信方必须把自己早 已随机生成的公钥送给发信方，而自己保留私钥。由于不对称算法拥有两个密钥，因而特 别适用于分布式系统中的数据加密。广泛应用的不对称加密算法有RSA 算法和美国国家标 准局提出的 DSA 。以不对称加密算法为基础的加密技术应用非常广泛。

(3)不可逆加密算法

不可逆加密算法的特征是加密过程中不使用密钥，输入明文后由系统直接经过加密算 法处理成密文，这种加密后的数据是无法被解密的，只有重新输入明文，并再次经过同样 不可逆的加密算法处理，得到相同的加密密文并被系统重新识别后，才能真正解密。显然， 在这类加密过程中，加密是自己，解密还得是自己，而解密就是重新加一次密，所应用的 密码也就是输入的明文。不可逆加密算法不存在密钥保管和分发问题，非常适合在分布式





**大数据基础与应用**

网络系统上使用，但因加密计算复杂，工作量相当繁重，通常只在数据量有限的情形下使 用，如广泛应用在计算机系统中的口令加密，利用的就是不可逆加密算法。随着计算机系 统性能的不断提高，不可逆加密的应用领域逐渐增大。在计算机网络中应用较多不可逆加 密算法的有 RSA 公司发明的 MD5 算法和由美国国家标准局建议的不可逆加密标准 SHS (Secure Hash Standard,安全杂乱信息标准)等。不可逆加密算法也可以应用于数据加密、 身份认证和数据安全传输。

**5.** **传输安全**

数据传输加密技术的作用是以防止通信线路上的窃听、泄露、篡改和破坏。数据传输 的完整性通常通过数字签名的方式来实现，即数据的发送方在发送数据的同时利用单向的 不可逆加密算法 Hash 函数或者其他信息文摘算法计算出所传输数据的消息文摘，并把该消 息文摘作为数字签名随数据一同发送。接收方在收到数据的同时也收到该数据的数字签名， 接收方使用相同的算法计算出接收到的数据的数字签名，并把该数字签名和接收到的数字 签名进行比较，如果二者相同，则说明数据在传输过程中没有被修改，数据完整性得到了 保证。

**6.** **身份认证**

身份认证指的是用户身份的确认技术，它是网络安全的第一道防线，也是最重要的一 道防线。由于网上的通信双方互不见面，必须在交易时(即交换敏感信息时)确认对方的 真实身份。身份认证要求参与安全通信的双方在进行安全通信之前，必须互相鉴别对方的 身份。保护数据不仅仅是要使数据正确、长久地存在，更重要的是，要使不该看到数据的 人看不到。为了完成上述功能，就必须依靠身份认证技术来给数据加上一把锁。数据存在 的价值就是需要被合理访问，所以，建立信息安全体系的目的应该是保证系统中的数据只 能被有权限的人访问，没有经授权的人则无法访问到数据。如果没有有效的身份认证手段， 访问者的身份就很容易被伪造，使得没有经授权的人仿冒有权限人的身份，这样，任何安 全防范体系就都形同虚设，所有安全投入就被无情地浪费了。

在企业管理系统中，身份认证技术可以密切结合企业的业务流程，阻止对重要资源的 非法访问。身份认证技术也可以解决访问者的物理身份和数字身份的一致性问题，给其他 安全技术提供权限管理的依据，身份认证是整个信息安全体系的基础。

在公共网络上的认证，从安全角度分为两类： 一类是请求认证者的秘密信息(口令) 在网上传送的口令认证方式；另一类是使用不对称加密算法，而不需要在网上传送秘密信 息的认证方式，这类认证方式中包括数字签名认证方式。

**7.** **口令认证方式**

口令认证的前提是请求认证者必须具有一个 ID, 该 ID 必须在认证者的用户数据库 (该数据库必须包括ID 和口令)中是唯一的。同时为了保证认证的有效性，必须考虑到 以下问题：

● 请求认证者的口令必须是安全的。





第14章 大数据安全

● 在传输过程中，口令不能被窃看、替换。

● 请求认证者在向认证者请求认证前，必须确认认证者的真实身份，否则将把口令 发给冒充的认证者。

口令认证方式还有一个最大的安全问题就是系统的管理员通常都能得到所有用户的口 令。因此，为了避免这样的安全隐患，通常情况下在数据库中保存的是口令的 Hash 值，通 过验证 Hash 值的方法来认证身份。

**8.** **企业安全**

关键数据是企业正常运作的基础。 一旦遭遇数据灾难，那么企业整体工作将陷入瘫痪， 带来难以估量的损失。

(1)远程备份与手工备份

从数据安全解决方案的角度来看，无论是远程备份还是手工备份，都面临着一个巨大 工作量的问题。针对这种情况，越来越多的企业用户选择了固定数据恢复服务商。从根本 意义来看，数据恢复与备份技术本身并不冲突，前者作为后者的补救手段。从企业用户角 度来看，远程备份将是很好的解决方案式备份，但是依然不可彻底摆脱对紧急性数据恢复 服务的需求。

(2)涉密资质成为筛选标准

在选择数据恢复服务商的同时，企业往往非常重视服务商的资质，包括技术实力与信 息保密能力，而大型企业甚至要求服务商能够在全国各地提供周到的本地化服务。对于企 业用户而言，选择固定的数据恢复服务商可以降低整体成本，而且这样也能确保恢复过程 中涉密数据不被外泄，同时整体恢复成功率也能有一定的保证。

(3)保护关键的业务数据方法

① 备份关键的数据。备份数据就是在其他介质上保存数据的副本。有两种基本的备份 方法：完整备份和增量备份。完整备份将把所选的数据完整地复制到其他介质。增量备份 仅备份上次完整备份以来添加或更改的数据。通过增量备份扩充完整备份通常较快且占用 较少的存储空间。可以考虑每周进行一次完整备份，然后每天进行增量备份。但是，如果 要在崩溃后恢复数据，则要花费较长的时间，‘因为首先必须要恢复完整备份，然后才恢复 每个增量备份。如果对此感到担忧，则可以采取另一种方案，每天晚上进行完整备份；只 需使备份在下班后自动运行即可。

② 建立权限。操作系统和服务器都可对由于员工的活动所造成的数据丢失提供保护， 可以根据用户在组织内的角色和职责而为其分配不同级别的权限。不应为所有用户提供“管 理员”访问权，这并不是维护安全环境的最佳做法，而是应制定“赋予最低权限”策略， 把服务器配置为赋予各个用户仅能使用特定的程序并明确定义用户权限。

③ 对敏感数据加密。对数据加密意味着把其转换为一种可以伪装数据的格式。 加密 用于在网络间存储或移动数据时确保其机密性和完整性。利用工具对加密文件进行解密的 授权用户可以访问这些文件。加密对其他访问控制方法是一种补充，且对容易被盗的计算 机(如便携式计算机)上的数据或网络上共享的文件提供多一层保护。Windows XP 和 Windows Small Business Server 2003 支持加密文件系统对文件和文件夹加密。



**大数据基础与应用**

把这三种方法结合起来，应该可以为大多数企业提供保证数据安全所需的保护级别。

**14.2** **大数据安全的内容**

大数据安全也如大数据名词一样，包括了两个含义， 一个含义是如何保障大数据计算 过程、数据形态、应用价值的安全；另一个含义是将大数据用于安全，也就是利用大数据 相关的技术，提升安全的能力和安全效果。前者是指如何保证大数据的安全，后者是指如 何用大数据来解决安全问题。

大数据的出现，对数据存储的物理安全性要求将越来越高，从而对数据的多副本与容 错机制提出了更高的要求。由于网络和数字化环境更容易使得犯罪分子获得个人的信息， 也有了更多不易被追踪和防范的犯罪手段，甚至出现更为高明的骗局。大数据本身、大数 据处理过程、大数据处理结果都有可能受到网络犯罪的攻击。大数据安全不仅考虑网络层 次的安全，而且还应考虑内部操作人员的安全防范和审计。

**14.2.1** **大数据的不安全因素**

对于大数据，存在下列不安全因素。

**1.** **大数据成为网络攻击的显著目标**

大数据的特点是数据规模大，达到 PB 级，而且复杂，并存在更敏感的数据，吸引了 更多的潜在攻击者。

数据的大量汇集，致使得黑客成功攻击一次就可获得更多数据，因此，降低了黑客的 进攻成本，增加了收益效率。

**2.** **大数据加大了隐私泄露风险**

大量数据的汇集加大了用户隐私泄露的风险。

① 数据集中存储增加了泄露风险。

② 一些敏感数据的所有权和使用权并没有明确界定，很多大数据的分析都没有考虑到 其中涉及的个体隐私问题。

**3.** **大数据威胁现有的存储和安防措施**

大数据集中存储的后果是导致多种类型数据存储在一起，安全管理不合规格。大数据 的大小也影响到安全控制措施的正确运行。安全防护手段的更新升级速度跟不上数据量非 线性增长的速度，进而暴露了大数据安全防护的漏洞。

**4.** **大数据技术成为黑客的攻击手段**

在利用大数据挖掘和大数据分析等大数据技术获取价值的同时，黑客也在利用这些大 数据技术发起攻击。黑客最大限度地收集更多有用信息，例如社交网络、邮件、微博、电





**第14章大数据安全**

子商务、电话和家庭住址等信息，大数据分析使黑客的攻击更加精准，大数据也为黑客发 起攻击提供了更多机会。

**5.** **大数据成为可持续攻击的载体**

传统的检测是基于单个时间点进行的基于威胁特征的实时匹配检测，而可持续攻击是 一个实施过程，无法被实时检测。此外，大数据的价值密度低，使得安全分析工具很难聚 焦在价值点上，黑客可以将攻击隐藏在大数据中，对安全服务提供商的分析制造了很大困 难。黑客设置的任何一个误导安全目标信息提取和检索的攻击，都将导致安全监测偏离方向。

**14.2.2** **大数据安全的关键问题**

**1.** **网络安全**

随着在网上进行越来越多的交易、对话和互动，使得网络犯罪分子比以往任何时候都 要猖獗。网络犯罪分子组织得更好、更专业，并具备有力的工具和能力，以针对确定的目 标进行攻击。这对企业造成声誉受损，甚至财政破产。网络弹性和防备战略对于企业大数 据至关重要。

**2.** **云中的数据**

由于企业迅速采用和实施新技术，例如云服务。所以经常面临大数据的存储和处理的 需求。而这其中包含了不可预见的风险和意想不到的后果。在云中的大数据对于网络犯罪 分子来说，是一个极具吸引力的攻击目标。这就对企业提出了必须构建安全的云的需求。

**3.** **个人设备安全管理**

大数据的出现扩大了移动设备使用范围，企业面临的是员工在工作场所使用个人设备 的安全管理挑战，因此必须平衡安全与生产力的需要。由于员工智能分析和浏览网页详情 混合了家庭和工作数据。企业应当确保其员工遵守个人设备相关的使用规则，并在符合其 既定的安全政策下管理移动设备。

**4.** **相互关联的供应链**

企业是复杂的、全球性的和相互依存的供应链的一个环节，而且是最薄弱的环节。信 息通过简单平凡的数据供应链结合起来，包括从贸易或商业秘密到知识产权的一系列信息， 如果损失就可能导致企业声誉受损，受到财务或法律的惩罚。信息安全协调在业务关系中 起着相当重要的作用。

**5.** **数据保密**

大数据产生、存储和分析过程中，数据保密问题将成为一个更大的问题。企业必须尽 快开始规划新的数据保护方法，同时监测进一步的立法和监管的发展。数据聚合和大数据 分析是保证企业营销情报的宝库，能够在针对客户情况的基础上，结合过去的采购模式和 以前的个人喜好进行销售，这是营销的法宝。但企业领导人应了解申请多个司法管辖区的





**大数据基础与应用**

法律和其他限制。企业还应该实现数据隐私最佳分析程序，建立相关透明度和问责制，但 不要忽视大数据、流程和技术的作用。

**14.2.3** **大数据安全措施**

**1.** **基础设施支持**

为了创建支持大数据环境的基础设施，需要一个安全且高速的网络来收集很多安全系 统数据源，从而满足大数据的收集要求。基于大数据基础设施的虚拟化和分布式性质，可 将虚拟网络作为底层通信基础设施。此外，从承载大数据的角度来看，在数据中心和虚拟 设备之间使用VLAN 等技术作为虚拟主机内的网络。由于防火墙需要检查通过防火墙的每 个数据包，这已经成为大数据快速计算能力的瓶颈。因此，企业需要分离传统用户流量与 大数据安全数据的流量。确保只有受信任的服务器流量流经加密网络通道以及防火墙，这 个系统就能够以所需要的不受阻碍的速度进行通信。

**2.** **保护虚拟服务器**

保护虚拟服务器的最好方法是确保这些服务器按照 NIST 标准进行加强，卸载不必要 的服务(如 FTP 工具)以及确保有一个良好的补丁管理流程。鉴于这些服务器上的数据的 重要性，还需要为大数据中心部署备份服务。此外，这些备份也必须加密，无论是通过磁 带介质还是次级驱动器的备份，在很多时候，安全数据站点发生数据泄露事故都是因为备 份媒介的丢失或者被盗。另外，应该定时进行系统更新，同时，为了进行集中监控和控制， 还应该部署系统监视工具。

**3.** **整合现有工具和流程**

因为数据量非线性增长，绝大多数企业都没有专门的工具或流程来应对这种非线性增 长。也就是说，随着数据量的不断增长，传统工具已经不再像以前那么有用。为了确保大 数据安全仓库位于安全事件生态系统的顶端，还必须整合现有安全工具和流程。当然，这 些整合点应该平行于现有的链接，因为企业不能为了大数据的基础设施改组而放弃其安全 分析功能。对于一项新部署，最好的方法是尽量减少连接数量，通过连接企业或业务线的 SIEM 工具的输出到大数据安全仓库。由于这些数据已经被预处理，将允许企业开始测试其 分析算法与加工后的数据集。

**4.** **制订严格的培训计划**

由于大数据在一个新的不同的环境运行，还需要为安全办公人员定制一个培训计划。 培训计划应该着眼于新开发的分析和修复过程，因为安全大数据仓库将通过这些过程来标 记和报告不寻常的活动和网络流量。大数据生态系统的实际操作有着非常标准化的功能， 没有经授权的更改或者访问将很容易被发现。

数据安全问题涉及企业的很多重大利益，发展数据安全技术是面临的迫切要求，除了 上述内容以外，数据安全还涉及其他很多方面的技术与知识，例如黑客技术、防火墙技术、





**第14章** **大数据安全**

入侵检测技术、病毒防护技术、信息隐藏技术等。 一个完善的数据安全保障系统，应该根 据具体需求对上述安全技术进行取舍。

**14.3** **云** **安** **全**

云存储是大数据生态环境的重要一环，云存储的安全隐忧令用户难以放心地将隐私数 据和应用部署在云端。 一方面是由于新的技术和新的计算模型产生的安全需求难以满足， 另一方面来自云内、云外对用户隐私保护的威胁。

**14.3.1** **云计算中用户的安全需求**

**1.** **云计算的安全需求**

用户对于云计算的主要安全需求如下所述。

① 执行安全需求，即用户的任务能够在可信的执行环境中正确地完成。

② 数据安全需求，即用户的隐私数据不会被第三方以及恶意的云服务提供商所窃取。

③ 服务安全需求，即用户能够在任何时间、任何地点无缝地接入云服务，并且接入全 过程安全、可靠和可验证。

在上述需求中，由于云计算的外包特性，用户的数据全部存储于云端，所以数据安全 是核心。

**2.** **用户的数据安全**

用户的数据安全包含以下三个方面。

(1)隐私性

用户的数据是私密的，不能被其他公司或个人(包括云服务提供商)所窃取。 (2)完整性

用户的数据是完备的、可信的，不能遭遇丢失或者未经授权的修改。同时，用户对数 据的修改、删除等操作最终能够真正被执行。

(3)一致性

用户的数据是统一的，不同的授权用户对同一份数据的访问结果应当相同。

与传统体系中数据存储在公司内部的服务器或者个人电脑上不同，在云计算中数据存 储在云端服务器上，因此用户丧失了对隐私数据的物理保护能力。同时，用户需要通过互 联网传输数据，更增加了数据泄露的风险。除此之外，数据的完整性也是用户对数据安全 的重要需求。如何保障用户数据不被损毁、不受未授权修改以及所有用户操作被忠实地执 行是云安全的重要课题。云平台还需要保证用户数据的一致性，即多个用户所看到的保存 在云端的同一份数据应该完全相同。攻击者往往是通过数据的不一致性访问未授权的数据， 或者实施进一步的攻击。用户代码与执行的安全是云计算安全的另一个重要方面。很多恶 意攻击都是通过劫持受害人机器的控制流来达到恶意目的。





大数据基础与应用

**14.3.2** **威胁模型**

对云安全的威胁可以根据攻击者的来源划分为来自云内部的威胁和来自云外部的威胁。

**1.** **来自云内部的威胁**

来自云内部的威胁是指云平台的管理员甚至是云服务提供商本身可能窥探用户的隐 私。具体而言，云内部的威胁可以进一步细化为下述三个级别。

① 无特权的云内部威胁；

② 云平台管理员或者云服务提供商恶意的威胁；

③ 物理安全级别的威胁。

**2.** **来自云外部的威胁**

来自云外部的威胁包括现有的常规攻击以及由云计算特征产生的旁路攻击和对虚拟机 监控器的攻击等。来自云内部的威胁是云计算所特有的威胁，云计算用户的隐私数据和执 行代码都保存和运行在云服务提供商的机器中，这是产生云内部威胁的源泉。来自云外部 的威胁可以是来自互联网，虚拟化层攻击者可以通过虚拟化层的漏洞反向攻击并劫持虚拟 机监控器，从而窃取其他用户的隐私。

**14.3.3** **云安全的支撑技术**

由于云计算环境的动态性和复杂性，传统的安全手段不能满足云环境中用户的需求， 如互联网中广泛应用的防火墙和网关杀毒等。

**1.** **防火墙**

防火墙主要进行访问控制，防止恶意和未经授权的流量进入内网。网关杀毒产品主要 负责扫描网络流量和文件内容，查杀木马、病毒等恶意代码。由于多个不同用户的虚拟机 可能共享一台物理主机，传统的防火墙很难插入虚拟机间进行访问控制。又由于，在云环 境中虚拟机随时被创建和销毁，而且恶意的用户也可以租用虚拟机，在这种环境下，网络 隔离也成为一个巨大的困难问题。

**2.** **虚拟层的接入将造成新的安全隐患**

目前安全产品是基于传统软件栈设计和构建的，虚拟化层的加入将有可能造成新的安 全隐患产生，如虚拟机间的通信也许不能被监控软件察觉到。除此之外，互联网中现有的 安全问题将同样威胁到云计算平台，甚至通过虚拟化层的传播可能使一个漏洞产生更广泛 的影响。如果一台服务器遭受到了攻击，虚拟机监控器被攻击者劫持，那么在其上运行的 所有虚拟机中的数据和应用都将毫无安全性可言。

**3.** **云计算的新特征带来了新问题**

除了传统的安全问题，云计算的新特性也带来了新的问题和挑战，例如利用共享资源





**第14章大数据安全**

的旁路攻击、对于虚拟化层的攻击、数据应用迁移中的安全性的保护、动态复杂云环境中 的可信问题等。

**14.3.4** **用户数据隐私保护**

在来自互联网的威胁中，由于云计算的规模特性使得传统的攻击手段造成的后果可能 更为严重。例如在云环境中， 一旦云平台被攻破，其上所有的虚拟机都将丧失保护。另外， 由于云计算平台的开放性，运行在其上的所有软件都可能成为攻击的目标，使得攻击面也 随之增大。由于虚拟机监控器及相应的软件的复杂度和代码行数的不断增加，虚拟化层的 漏洞也在不断增多。如何保护隐私数据不受到云内外攻击的威胁，如何保护用户数据能够 不被非法篡改和窃取，是当前研究的重点问题。

对于隐私数据的保护存在如下问题。

① 云计算中敏感数据和非敏感数据混杂存储，不同虚拟机共享同样的存储介质，难以 进行权限控制。

② 云服务提供商监守自盗，盗窃用户的隐私数据。

③ 用户外包数据的控制权，隐私数据的完整性难以保证。

用户的隐私数据可以细化为静态数据和动态数据两种。静态数据是指用户的文档、报 表、资料等不参与计算的隐私信息。动态数据则是指需要动态验证或参与计算的数据。面 对用户静态数据隐私泄露问题，使用数据加密技术是一个简捷而有效的方法。用户可以使 用加密机制对数据进行加密，并将加密后的数据保存在云端。另一个解决方法是在云端使 用加密文件系统，这样可以保证在磁盘中的文件均以密文形式保存。然而这种先加密再存 储的方法只适用于静态的数据，不适用于参与运算的动态的数据，因为动态数据需要在CPU 和内存中以明文形式存在。迄今为止，对于用户动态数据隐私保护还没有一种彻底的解决 方案。

**14.3.5** **云计算执行环境的可信性**

利用公有云能够通过服务托管和服务整合，削减在计算资源方面的成本并提供强大的 灵活性和可伸缩性。然而基础设施公有化带来租户对于自身任务应用和数据安全的不可控， 引发了租户对于云平台的信任问题。多层次的服务模式(如基础设施即服务、平台即服务 和软件即服务)之间的信任，租户执行环境中动态定制的服务组件之间的信任都是云计算 平台亟待解决的问题。除此之外，在云平台内部，随着功能增多与代码膨胀，云平台的可 信性将暴露出更多的安全漏洞。执行环境的可信基是指云端计算环境忠实执行其设定的功 能，并且不执行任何没有设定的功能。在可信计算中，用户可以使用可信平台模块TPM 保 证启动环境的可信性。TPM 是可信计算组织 (Trusted Computing Group,TCG) 为了构建 可信的运行环境提出的一种具有安全存储和加密功能的可信平台模块。TPM 使用信任链传 递的方式将启动代码的哈希值存放在芯片中，用户可以通过远程验证比较哈希值确定远端 代码是否可信。然而，与TPM 类似的基于哈希的认证很难验证依赖输入或外部数据的程序，





**大数据基础与应用**

如 Web 服务器等，所以相关技术只能保证启动时安全而不能保证运行时安全。经验表明， 越小规模的程序具有越强的可信性。因此，可以通过减小规模来减少系统中的漏洞，从而 增强执行环境的可信性。

系统也可以通过增加基于凭证的认证机制保证程序运行环境的可信性。TCG Linux 就 是将认证能力加入Linux 中形成的系统，使得传统的 Linux 系统更加可信。然而由于 Linux 系统规模过大，并且缺少强隔离机制， TCG Linux 很难真正保证攻击者无法绕过验证机制。

**14.3.6** **资源共享问题**

云计算的重要特征之一是云中的多用户共享同一物理设备。资源共享能够有效地利用 闲置的硬件计算、存储能力，为用户提供更强的可拓展性和灵活性，然而，资源共享也可 能会给用户带来隐私泄露的问题。旁路攻击就是资源共享所导致，它是指使用来自物理实 现等非直接传输途径的信息，而不是系统本身逻辑漏洞进行的攻击。由于旁路攻击能够绕 开系统本身的防御逻辑，所以防御困难。目前旁路攻击主要集中于共享CPU 缓存的环境中， 内存的去冗余机制也可能使攻击者窥探到用户运行程序。资源共享的另一个问题是不同用 户间的共享问题。在一些情况下，如果在云中有一个恶意用户的存在，那么有可能会导致 许多无辜用户的服务不可用，从而变相地攻击了这些用户的服务可用性。

小结

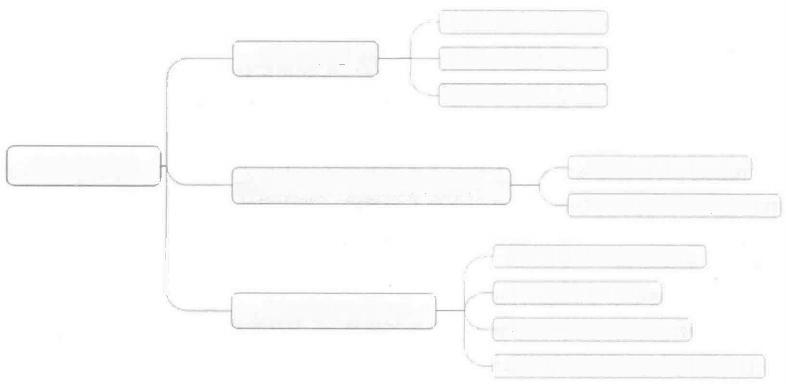
云计算安全面临着数据隐私泄露、执行环境不可信、旁路攻击等多种问题。现有的方 法利用加密技术、隔离技术、控制流完整性、可信计算等手段保护用户的数据和执行的隐 私性和完整性，取得了一定的效果。但是，现有的安全技术应用于云端将会面临大数据、 高并发访问、云服务所要求的可用性、软件兼容性等问题，保障云服务的效率与质量的同 时增强整个系统的安全性将是重要的研究课题。



**第15章**

本章主要内容

**大数据机器学习**



机器学习的产生与发展

机器学习类型

知识表示形式

大数据机器学习的特点

大数据机器学习的特点与评测指标

大数据机器学习的评测指标

基于大数据的空气质量推断

人与建筑的关系分析

大数据机器学习的应用

针对全球问题的预测模型

全球地表覆盖制图可视化与数据分析

大数据机器学习

机器学习概述

机器学习是一门迅速发展的多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、 算法复杂度理论等多门学科及领域。机器学习在许多应用领域已得到成功的应用，已成为 计算机科学的研究热点之一。机器学习方法已在数据挖掘、计算机视觉、自然语言处理、 生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA 序列测序、 语音和手写识别、战略游戏和机器人运用等领域广泛应用。随着大数据的出现，大数据机 器学习的方法备受重视，并具有广泛的应用前景。

**15.1** **机器学习概述**

**15.1.1** **机器学习的产生与发展**

**1.** **机器学习发展的四个时期**

机器学习是人工智能研究中一个分支，产生于20世纪50年代。它的发展过程大体上 可分为如下四个时期。

从20世纪50年代中叶至60年代中叶，称为机器学习的发展时期。

从20世纪60年代中叶至70年代中叶，称为机器学习的冷静时期。

从20世纪70年代中叶至80年代中叶，称为复兴时期。



**大数据基础与应用**

1986年到现在，机器学习进入新的时期，发展迅速，主要的表现如下所述。 (1)机器学习已成为新的边缘学科

机器学习综合应用心理学、生物学和神经生理学以及数学、自动化和计算机科学形成 了机器学习理论基础，并已在高校形成了一门独立的课程。

(2)混合式的集成学习系统研究方兴未艾

结合各种学习方法，取长补短的多种形式的集成学习系统研究正在兴起。特别是连接 学习与符号学习的耦合可以更好地解决连续性信号处理中知识与技能的获取与求精问题而 受到重视。

(3)机器学习与人工智能统一性观点正在形成

学习与问题求解结合进行，知识表达便于学习的观点产生了通用智能系统 SOAR 的组 块学习。类比学习与问题求解结合的基于案例方法已成为经验学习的重要方向。

(4)各种机器学习方法的应用范围不断扩大

归纳学习的知识获取工具已在诊断分类型专家系统中广泛使用。连接学习在声图文识 别中占优势。分析学习已用于设计综合型专家系统。遗传算法与强化学习在工程控制中有 较好的应用前景。与符号系统耦合的神经网络连接学习将在企业的智能管理与智能机器人 运动规划中发挥作用。

(5)大数据促进了机器学习的发展

大数据技术的出现对各信息领域提出了新的挑战，尤其是推动了机器学习方法的研究， 提出了基于大数据的机器学习模式。

**2.** **机器学习的概念**

学习是人类具有的一种重要智能行为。著名的人工智能学者西蒙教授认为，学习就是 系统在不断重复的工作中对本身能力的增强或者改进，使得系统在下一次执行同样任务或 类似任务时，将做得更好或效率更高。

1959年美国的塞缪尔 (Samuel) 设计了一个具有学习能力的下棋程序，这个程序可以 在不断的对弈中改善自己的棋艺。4年之后，这个程序战胜了设计者本人。又过了3年，这 个程序战胜了美国一个头衔保持8年之久的冠军，展示了机器学习的能力。

机器学习是研究如何使用机器来模拟人类学习活动的一门学科。更为严格地说，机器 学习是一门研究机器获取新知识和新技能，并识别现有知识的学科。这里所说的机器是计 算机，现在是电子计算机，以后还可能是中子计算机、生物计算机、光子计算机或神经计 算机等。

机器学习的过程是：将一些已知的并已被成功解决的问题作为范例输入计算机，机器 通过学习范例，总结并归纳具有通用性的规则，然后，使用这些规则解决某一类的问题。 在此之后，神经网络技术的形成和发展，关注点转向知识工程。知识工程是对计算机输入 范例而生成规则，计算机是通过使用这些规则来解决某些问题，例如专家系统等。这种方 法投资大、效果不理想。20 世纪80年代又在新的神经网络理论的指导下，重新回到机器 学习的方法，并将其成果应用于处理大型商业数据库。它就是数据库中的知识发现KDD, 它泛指所有从源数据中发掘模式或联系的方法，并用 KDD 来描述整个数据发掘的过程，





**第15章大数据机器学习**

包括最开始的制定业务目标到最终的结果分析，而用数据挖掘来描述使用挖掘算法进行数 据挖掘的子过程。数据挖掘中的工作可以利用统计方法来完成，并认为将统计方法与数据 挖掘结合的策略是最好的策略。

机器学习的应用包括：通过分析以往销售数据来预测客户行为，人脸识别或语音识别， 优化机器人行为以便使用最少的资源来完成任务，以及从生物信息数据中提取知识等。

**3.** **机器学习理论及研究**

机器学习是关于理解与研究学习的内在机制、建立能够通过学习自动提高自身水平的 计算机程序的理论方法的学科。

(1)机器学习的理论

机器学习理论主要是设计和分析计算机自动学习的算法。机器学习算法是一类从数据 中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大 量的统计学理论，机器学习与统计推断学联系尤为密切，也被称为统计学习理论。在算法 设计方面，机器学习理论关注可以实现的、行之有效的学习算法。很多推论问题属于无程 序可循，所以部分的机器学习研究是开发容易处理的近似算法。

(2)机器学习的研究

机器学习的研究是根据生理学、认知科学等对人类学习机理的了解，建立人类学习过 程的计算模型或认识模型，发展各种学习理论和学习方法，研究通用的学习算法并进行理 论上的分析，建立面向任务的具有特定应用的学习系统。在整个机器学习的发展历程中， 一直有两大研究方向。 一是研究学习机制，注重探索和模拟人的学习机制；二是研究如何 有效利用信息，注重从海量数据中获取隐藏的、有效的和可理解的知识。

机器学习领域的研究主要围绕以下三个方面进行。

① 面向任务的研究：研究和分析改进一组预定任务的执行性能的学习系统。 ② 认知模型研究：分析人类学习过程并进行计算机模拟。

③ 理论分析：从理论上探索各种可能的学习方法和独立于应用领域的算法。

机器学习是继专家系统之后人工智能应用的又一重要研究领域，也是人工智能和神经 计算的核心研究课题之一。现有的系统学习能力有限，因而不能满足科技和生产提出的新 要求。

学习是一项复杂的智能活动，学习过程与推理过程紧密相连，按照学习中使用推理的 多少，机器学习所采用的策略可分为机械学习、传授学习、类比学习和事例学习。

**4.** **机器学习系统的结构**

机器学习系统是指能够在一定程度上实现机器学习的系统。机器学习系统能够从某个 过程或环境的未知特征中学到有关信息，并且能够把学到的信息用于未来的估计、分类、 决策或控制，以便改进系统的性能。也就是说，在与环境相互作用时，能够利用过去与环 境作用时得到的信息，提高其性能。如图15-1 所示为机器学习系统的基本结构。

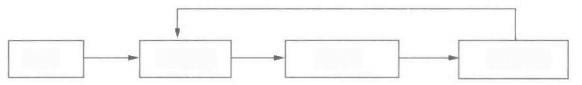
在图15-1 中，环境为系统的学习部分提供信息，学习部分利用这些信息修改知识库， 以增进系统执行部分完成任务的效能，执行部分根据知识库完成任务，同时把获得的信息





**大数据基础与应用**

反馈给学习部分。在具体的应用中，环境、知识库和执行部分决定了具体的工作内容，学 习部分所需要解决的问题完全由上述三部分确定。下面分别叙述这三部分对设计学习系统 的影响。



环境 学习部分 知识库 执行部分

图15-1 机器学习系统的基本结构

影响学习系统设计的最重要的因素是环境向系统提供的信息，更具体地说是信息的质 量。知识库中存放的是指导执行部分动作的一般原则，但环境向学习系统提供的信息各种 各样。如果信息的质量比较高，与一般原则的差别比较小，则学习部分比较容易处理。如 果向学习系统提供的是杂乱无章的指导执行具体动作的具体信息，则学习系统需要在获得 足够数据之后，删除不必要的细节，进行总结推广，形成指导动作的一般原则，放入知识 库，这样学习部分的任务就比较繁重，设计起来也较为困难。因为学习系统获得的信息往 往是不完全的，所以学习系统所进行的推理并不完全可靠，总结出来的规则不能保证完全 正确。这要通过执行效果加以检验。正确的规则能使系统的效能提高，应予保留，不正确 的规则应予修改或从知识库中删除。

知识库是学习系统存储知识的集合，是影响学习系统设计的第二个因素。知识的表示 有多种形式，比如特征向量、 一阶逻辑语句、产生式规则、语义网络和框架等。在选择 表示方式时要兼顾以下四个方面。

● 表达能力强；

● 易于推理；

● 容易修改知识库；

● 知识表示易于扩展。

需要说明的是，知识库学习系统不能在全然没有任何知识的情况下凭空获取知识，每 一个学习系统都要求具有某些知识理解环境提供的信息，分析比较，做出假设，检验并修 改这些假设。因此，更确切地说，学习系统是对现有知识的扩展和改进。执行部分是整个 学习系统的核心，因为执行部分的动作就是学习部分力求改进的动作。同执行部分有关的 问题是复杂性、反馈和透明性。

**15.1.2** **机器学习类型**

机器学习类型繁多，基于不同的分类原则，分类如下。

**1.** **基于学习策略的学习**

在一个学习系统中，由环境部分(如书本或教师)提供信息，学习部分则实现信息转 换，从中获取有用的信息，并使用能够理解的形式记忆下来。在学习过程中，学生(学习 部分)使用的推理越少，他对教师(环境)的依赖就越大，教师的负担也就越重。学习策 略是指学习过程中系统所采用的推理策略，也就是说，基于学习策略的分类就是根据学生





**第15章大数据机器学习**

实现信息转换所需的推理多少和难易程度的分类，基于从简单到复杂，从少到多的次序可 以分为以下六种基本类型。

(1)机械学习

学习者无须任何推理或其他的知识转换，直接吸取环境所提供的信息。机械学习主要 考虑的是如何对存储的知识编制索引并加以利用，学习方法是直接通过事先构造的程序来 学习，学习者不做任何工作，或者是通过直接接收既定的事实和数据进行学习，对输入信 息不做任何的推理。

(2)示教学习

学生从环境获取信息，把知识转换成内部可使用的表示形式，并将新的知识和原有知 识有机地结合为一体。所以要求学生有一定程度的推理能力，但环境仍要做大量的工作。 教师以某种形式提出和组织知识，以使学生拥有的知识可以不断地增加。这种学习方法和 人类社会的学校教学方式相似，学习的任务就是建立一个系统，使它能接受教导和建议， 并有效地存储和应用学到的知识。

(3)演绎学习

学生所用的推理形式为演绎推理。推理从公理出发，经过逻辑变换推导出结论。这种 推理是保真变换和从一般到特殊的过程，使学生在推理过程中可以获取有用的知识。演绎 推理是归纳推理的逆过程。

(4)类比学习

类比学习是指利用源域和目标域中的知识相似性，通过类比转变为适应于新的领域的 知识，来完成原先没有设计的相类似的功能。类比学习需要比上述三种学习方式更多的推 理。它一般要求先从知识源(源域)中检索出可用的知识，再将其转换成新的形式，用到 新的状况(目标域)中去，类比学习是科学发现的常用方法。

(5)解释学习

学生根据教师提供的目标概念、该概念的一个例子、领域理论及可操作准则，首先构 造一个解释来说明为什么该例子满足目标概念，然后将解释推广为目标概念的一个满足可 操作准则的充分条件。

(6)归纳学习

归纳学习是由教师或环境提供某概念的一些实例或反例，让学生通过归纳推理得出该 概念的一般描述。归纳学习的推理工作量远多于示教学习和演绎学习，因为环境并不提供 一般性概念描述(如公理)。从某种程度上说，归纳学习的推理量比类比学习大，因为没有 一个类似的概念可以作为源概念加以取用。归纳学习是最基本的、较为成熟的学习方法， 在人工智能领域中已经得到广泛的研究和应用。

**2.** **基于应用领域的学习**

从机器学习的执行部分所反映的任务类型上看，大部分的应用研究领域基本上集中于 分类和问题求解两个范畴。

① 分类任务要求系统依据已知的分类知识对输入的未知模式(该模式的描述)做分析， 以确定输入模式的类属。相应的学习目标就是学习用于分类的准则，如分类规则等。





**大数据基础与应用**

② 问题求解任务要求对于给定的目标状态，寻找一个将当前状态转换为目标状态的动 作序列，机器学习在这一领域的研究工作大部分集中于通过学习来获取能提高问题求解效 率的知识，如搜索控制知识、启发式知识等。

**3.** **基于综合因素的学习**

综合考虑各种因素，可将机器学习方法区分为以下六类。

(1)经验归纳学习

经验归纳学习采用一些数据密集的经验方法对例子进行归纳学习，其例子和学习结果 一般都采用属性、谓词、关系等符号表示。它相当于基于学习策略分类中的归纳学习，但 除去联结学习、遗传算法、加强学习的部分。

(2)分析学习

分析学习方法是从一个或少数几个实例出发，运用领域知识进行分析。其主要特征为： 推理策略主要是演绎，而非归纳。使用过去的问题求解经验(实例)指导新的问题求解， 或产生能更有效地运用领域知识的搜索控制规则。

分析学习的目标是改善系统的性能，而不是新的概念描述。分析学习包括应用解释学 习、演绎学习、多级结构组块以及宏操作学习等技术。

(3)类比学习

类比学习相当于基于学习策略分类中的类比学习。在这种学习类型中最具有代表性的 是基于范例的学习，或简称范例学习，它是通过与过去经历的具体事例做类比来学习。

(4)遗传学习

遗传算法模拟生物繁殖的突变、交换和达尔文的“适者生存，不适者淘汰”的自然选 择。它将问题可能的解编码为一个称为个体的向量，向量的每一个元素称为基因，并利用 目标函数对群体(个体的集合)中的每一个个体进行评价，根据评价值(适应度)对个体 进行选择、交换、变异等遗传操作，从而得到新的群体。遗传算法适用于非常复杂的环境， 例如，带有大量噪声和无关数据、事物不断更新、问题目标不能明显和精确地定义，以及 通过很长的执行过程才能确定当前行为的价值等。

(5)联结学习

典型的联结模型实现为人工神经网络，其由称为神经元的一些简单计算单元以及单元 间的加权联结组成。

(6)加强学习

加强学习是通过与环境的试探性交互来确定和优化动作的选择，进而实现序列决策任 务。在这种学习中，学习机制通过选择并执行动作，导致系统状态的变化，并获得强化信 息，从而实现与环境的交互。强化信息就是对系统行为的一种标量化的奖惩。系统学习的 目标是寻找一个合适的动作选择策略，即在任一给定的状态下选择动作的方法，使产生的 动作序列可获得某种最优的结果。

在综合因素学习类型中，经验归纳学习、遗传算法、联结学习和加强学习属于归纳学 习，其中经验归纳学习采用符号表示方式，而遗传算法、联结学习和加强学习则采用亚符 号表示方式，分析学习属于演绎学习。实际上，类比学习是归纳和演绎策略的综合。所以





**第15章大数据机器学习**

最基本的学习策略只有归纳和演绎。从学习内容的角度看，采用归纳策略的学习由于是对 输入进行归纳，所学习的知识显然超过原有系统知识库所能蕴含的范围，所学结果改变了 系统的知识演绎闭包，因而这种类型的学习又可称为知识级学习。而采用演绎策略的学习 尽管所学的知识能提高系统的效率，但仍能被原有系统的知识库所蕴含，即所学的知识未 能改变系统的演绎闭包，因而这种类型的学习又被称为符号级学习。

**15.1.3** **知识表示形式**

学习系统获取的知识主要有下述表示形式。

**1.** **代数表达式参数**

学习的目标可以通过调节一个固定函数形式的代数表达式参数或系数来达到一个理想 的性能，因此，代数表达式参数可以作为获取的知识表示形式。

**2.** **决策树**

用决策树来划分物体的类属，树中每一内部节点对应一个物体属性，而每一边对应于 这些属性的可选值，树的叶节点则对应于物体的每个基本分类。

**3.** **形式文法**

在识别一个特定语言的学习中，通过对该语言的一系列表达式进行归纳，形成该语言 的形式文法。

**4.** **产生式规则**

产生式规则表示为条件一动作对，已被极为广泛地使用。学习系统中的学习行为主 要是生成、泛化、特化或合成产生式规则。

**5.** **形式逻辑表达式**

形式逻辑表达式的基本成分是命题、谓词、变量、约束变量范围的语句以及嵌入的逻 辑表达式。

**6.图**

有的系统采用图匹配和图转换方案来有效地比较和索引知识。

**7.** **框架和模式**

每个框架包含一组槽，用于描述事物(概念和个体)的各个方面。

**8.** **程序与过程编码**

获取这种形式的知识的目的在于取得一种能实现特定过程的能力，而不是为了推断该 过程的内部结构。





**大数据基础与应用**

**9.** **神经网络**

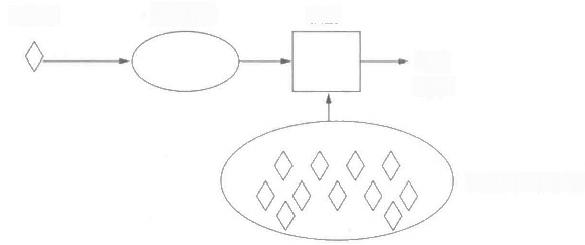
这主要用在联结学习中。学习所获取的知识，最后存于一个神经网络中。

**10.** **多种表示形式的组合**

一个学习系统中获取的知识有时需要综合应用上述几种知识表示形式。根据表示的精 细程度，可将知识表示形式分为两大类：泛化程度高的粗粒度符号表示、泛化程度低的精 粒度亚符号表示，如决策树、形式文法、产生式规则、形式逻辑表达式、框架和模式等属 于符号表示类，而代数表达式参数、图、神经网络等则属亚符号表示类。

**15.2** **大数据机器学习的特点与评测指标**

传统的机器学习方法首先找到一个新的样本空间，从中提取一些特征信息，然后以少 量数据作为训练样本建立模型，再将测试数据与模型进行匹配，达到对数据进行分类的目 的。上述过程如图15-2所示。



特征表示

结果

是/否

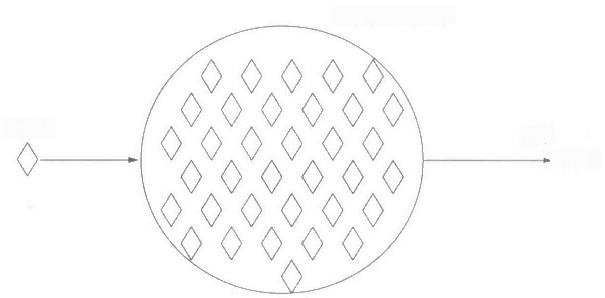
训练样本(小数据)

新样本

模型

图15-2 传统的机器学习方法

大数据改变了机器学习的方法，主要是取消了特征提取和建模的过程，大数据机器学 习的方法如图15-3所示。

已知样本(大数据)

结果

新样本

→是/否

图15-3 大数据机器学习方法





**第15章** **大数据机器学习**

在大数据机器学习中，数据量越大，就越不容易建立模型，这是因为样本空间本身就 将模型包含在内了。但是，在使用大数据进行机器学习时，应需要考虑下述问题。

● 数据的覆盖度：对于所有或大部分事件是否有足够的样本来覆盖。

● 数据的精度：对于高频事件，是否有足够的样本来提升它的精度。

大数据技术具有从各种类型的数据(包括结构化、半结构化和非结构化数据)中快速 获得有价值信息的能力。传统数据分析技术注重利用预先设定的统计方法对数据进行分析， 发现数据的价值。与传统数据分析相比，大数据技术的一个核心目标是要从体量巨大、结 构繁多的数据中挖掘出隐含的规律，进而获得最大的价值。从大量结构繁多的数据中挖掘 隐含规律必须与机器学习相结合，由计算机代替人去挖掘信息，获取知识。大数据技术的 目标实现与机器学习的发展密切相关。

大数据的发展从研究方向、评测指标以及关键技术等方面对机器学习提出了新的需求。 学习机制的研究是机器学习产生的源泉，但随着各行业对大数据分析需求的持续增长，通 过机器学习高效地获取知识，已逐渐成为机器学习技术发展的主要推动力。大数据机器学 习更强调学习本身是手段，机器学习成为一种支持技术和服务技术，如何基于机器学习对 复杂多样的数据进行深层次的分析，更高效地利用信息成为当前机器学习研究的主要方向。 机器学习越来越面向智能数据分析的方向发展，并已成为智能数据分析技术的一个重要技 术。另外，在大数据时代，随着数据产生速度的持续加快，数据的体量有了前所未有的增 长，而需要分析的新的数据种类也在不断涌现，如文本的理解、文本情感的分析、图像的 检索理解、图形和网络数据的分析等。机器学习研究领域出现了很多新的研究方向，例如， 利用未标识数据的半监督学习、有效解决训练数据质量问题、提高学习结果的泛化能力的 集成学习、在不同的领域进行知识迁移的迁移学习等。该研究领域提出了很多新的机器学 习方法，例如，大数据机器学习的有效分析方法，大数据并行处理方法等。

**15.2.1** **大数据机器学习的特点**

**1.** **大量的数据实例**

在互联网和金融领域，训练实例的数量非常大，每天汇集了几十亿事件的数据集合， 这样的数据集可以轻易地达到几百 TB 。有效处理大数据集的比较好的方式是组合服务器 云。MapReduce 分布计算框架就是通过简单的可并行化的函数语言原语将编程框架、使用 高容量存储及执行平台的能力有效地组合在一起，使大数据的计算变得更加容易。

**2.** **输入高维度的数据**

机器学习的应用包括自然语言、图形或者视频，这些应用中的数据实例是由很多数量 的特征来表示的，远超目前常见的能够轻松处理的特征量级别。在特征空间进行并行计算 是可以将计算扩展得更丰富的表示方法，还可以加入一些在特征空间进行迭代运算的算法。

**3.** **模型和算法的复杂性**

由于现有的许多机器学习算法是基于内存的，大数据却无法装载进计算机内存，故现





**大数据基础与应用**

有的诸多算法不能处理大数据。如何提出新的机器学习算法以适应大数据处理的需求，是 大数据时代的研究热点方向之一。

一些高准确性的学习算法，基于复杂的非线性模型或者采用昂贵的计算子程序。在这 两种情况下，将计算分配到多个处理单元是大数据学习算法的关键点。某些领域的数据在 一些基本的特征上是非线性结构。这类算法的一个共同特征是计算复杂性，单台机器的学 习过程可能会非常慢，可采用并行多节点或者多核处理，提高在大数据中使用复杂算法和 模型的计算速度。

**4.** **计算时间的限制**

很多应用都需要进行实时预测，需要推理算法的并行化。决定系统计算时间的因素是 单任务的处理时间和时延。单任务的处理时间的缩短可以通过提高系统单机的处理能力和 吞吐量来解决。时延出现在绝大多数应用场合，任务由多个相互关联的进程组成，不同进 程的处理时间长短不一，任务整体的处理实际有待于各个进程的结果。

**5.** **预测级联**

很多现实中的问题，如物体追踪、话音识别以及机器翻译，都要求执行一系列互相依 赖的预测，形成预测级联。如果一个级联作为一个推理任务，就将有一个很大的联合输出 空间。由于计算复杂性的增加，通常将导致非常高的计算成本。在预测任务之间的互相依 赖性通常是通过对单个任务进行阶梯式的并行化以及自适应的任务管理来实现的。

**6.** **模型选择和参数扫描**

调整学习算法的超参数以及统计重要性评估要求多次执行学习和推理，这些过程属于 可并行化应用，本身就非常适合并发执行。在参数扫描中，学习算法在配置不同的同一个 数据集上运行多次，然后在一个验证集上进行评估。

**15.2.2** **大数据机器学习的评测指标**

大数据机器学习技术在以下方面具有较好的适应力，这也是大数据机器学习的评测 指标。

**1.** **泛化能力**

期望经样本训练的机器学习算法具有较强的泛化能力，也就是能对没有学习过的新输 入给出合理响应的能力，这是评估机器学习算法性能的最重要指标。机器学习基本的目标 是对训练数据中的实例进行泛化推广。这是由于在测试时要再次看到与样本同样的实例概 率较小。

**2.** **可理解性**

许多功能强大的机器学习算法基本上都是黑盒子，对用户而言，只能看到输出结果， 却不知道为什么是这样的结果。随着数据量的增加、问题复杂度的提高，人们在得到结果





第15章 大数据机器学习

的同时更加希望了解为什么得到这样的结果，为此，应进行去黑盒化的研究。

**3.** **速度**

在机器学习中，需要考虑训练速度和测试速度等参数，这两者是互相关联的。对训练 速度和测试速度都具有很好指标的机器学习算法的研究异常重要。

**4.** **数据利用能力**

收集的数据不仅包括有标识的数据，还有大量未标识的数据。在信息处理过程中使用 标识数据，不使用未标识数据，将影响到的模型的泛化能力。所以，研究并开发能够有效 利用所有数据的机器学习方法，具有非常重要的实际意义，大数据机器学习更注重全样本 和大样本的机器学习技术。

**5.** **代价敏感**

大数据分析的精髓就在于综合各种内部、外部数据对一个事物进行全面的刻画和解读， 涉及的因素很多。代价敏感的学习算法就是这方面的一个有效的解决方案，通过引入代价 信息来度量误判的程度，不同的代价参数代表不同的损失，最终的目标是最小化总的代价 而不是总的错误。

**6.** **知识的迁移性**

通过迁移学习，将已有的学习成果能不断积累并衍生引用到未知的领域。当然，与大 数据相关的机器学习研究领域还有很多，如大规模机器学习的算法并行化，这也是未来机 器学习的一个重点方向。目前，在计算机视觉、语音识别、模式识别、建模和预测等领域 取得了显著性的进步。这些进步应用了新的机器学习技术，创建了大数据分析的基本模式 算法。

大数据机器学习的不断进展可以解决更多的难题，对整个社会产生深远的影响。

**15.3** **大数据机器学习的应用**

这里介绍几种典型应用实例。

**15.3.1** **基于大数据的空气质量推断**

建设智慧城市与改善生活品质是社会的热门议题，它激发了人们对了解所在地区空气 质量的渴望与需求。如果能够更准确地了解我们生活环境的空气质量，我们就能做出更明 智且更有利于健康的决策。例如何时何地最适合户外运动，或者何时应戴上口罩、关上窗 户。不同地点的空气质量差异很大，而且其成因也十分复杂，交通流量及土地使用情况等 都会对其产生影响。目前人们只能借助监测站才能准确判断某个地点的空气质量，然而监 测站却并非随处可见。为了应对这一挑战，可以根据现有监测站所提供的空气质量数据以



**大数据基础与应用**

及城市里的其他多种数据来源(包括气象情况、交通流量、人员流动趋向、路网结构、人 口集中点等),运用数据挖掘和机器学习技术，对大数据加以充分利用，并在监测信息和对 应结果之间建立一个隐式映射，从而可以实时推断出包含细颗粒物信息的城市空气质量数 据，帮助人们更方便地获取那些能够提高生活品质的实用信息。

**15.3.2** **人与建筑的关系分析**

建筑无处不在并影响着人类的生态印迹、身心健康和审美观念。为了使城市更加宜 居，近年来建筑界和学术界在了解建筑物的本质方面都取得了显著进步，并且在不断进 行优化。人类在影响建筑物的同时，也反过来受到建筑的影响，以人为中心对建筑物进 行数据分析非常关键。人与建筑分析是一种从居住者个体的角度进行个性化数据分析的 方法。人与建筑分析通过分析居住者的行为方式，然后帮助建筑物适应居住者的行为偏 好。借助低成本和定制化的传感器，可以将居住者行为模式与建筑物的相关数据分析联 系起来，即可对个人进行能耗追踪和行为分析，对电梯等楼宇服务进行优化，使我们的 建筑更智能、更宜居。

**15.3.3** **针对全球问题的预测模型**

在21世纪，在全球范围内，人类共同面临的挑战就是气候变化、森林过度砍伐、农业 生产力落后、人口增长过快、人口迁徙和水资源安全等。为了应对这些挑战，人们迫切需 要适用于关键性全球环境系统的预测模型，探究各种方法，解决预测这一难题。利用云计 算系统，对环境数据进行存储和分析，再运用机器学习技术建立预测模型，将其应用于预 测服务中，并可以根据需要在任何设备上运行。这些预测可以告诉我们何种严重的问题可 能会在何时何地发生，并推测出采取相应措施来减缓或解决这些问题的结果及影响。它能 够帮助用户实现数据的可视化，以环境信息对数据进行补充，定义复杂模型，使用贝叶斯 法确定参数，进行带有不确定性因素的预测，并且以完全透明且可重复的形式加以共享。 借助这一技术，预测将变得简单易行。

**15.3.4** **全球地表覆盖制图可视化与数据分析**

为导航或其他目的而绘制基础地图的日子早已成为过去。当前，正在发掘大数据资源， 并且以现有科研能力为基础应对新的挑战，开发出一种高空间分辨率全球地表覆盖测绘的 全新方法。通过它，有望大大改善地表覆盖测绘数据的可视化、分析、管理和传播。可以 制作出全球第一张30米分辨率的全球地表覆盖图。此外，这项技术还非常易于使用。它可 以让生态学家、地理学家和其他科学家在各自的研究领域(如气候变化、改善生态系统、 城市和农业规划、水文、地貌和大气模型、栖息地和生物多样性、碳循环以及公众健康等) 大受裨益。





第15章 大数据机器学习

小结

机器学习是一个重要领域，尤其近年来发展迅速，应用广泛。大数据的出现促进了机 器学习的发展，大数据机器学习可以解决以往很困难的问题或根本不能够解决的问题。本 章简单地介绍了有关大数据机器学习的内容，主要包括机器学习理论、机器学习类型、机 器学习系统的基本结构，尤其对大数据机器学习特点、评测标准和大数据机器学习的应用 做了进一步的介绍。从大数据机器学习过程可以看出，缺少了特征提取和建模的过程，但 是，这并不是说模型就没有使用价值了，这是由于数据永远是不够用的，以及由于人的主 观性，在样本空间中存在认为的偏见和噪声。另外，样本空间也不是一成不变，它随着时 间的推移不断变化，基于上述，需要采用一种模型与数据相结合的方法来提高机器学习的 效率。



**第16章**

本章主要内容

**大数据推荐技术**



推荐系统的产生与发展

推荐系统的概念

推荐系统架构

推荐系统类型

推荐系统的评判标准 推荐算法

推荐模式

下一代推荐系统

数据稀疏性问题

一大数据冷启动

多样性与精确性的两难命题

大数据处理与增量计算问题

推荐系统的脆弱性问题

大数据推荐技术的挑战

推荐系统效果评估

用户行为模式的挖掘和利用

用户界面与用户体验

多维数据的交叉利用

社会推荐

大数据人才推荐系统

推荐算法与推荐模式

大数据推荐技术

推荐技术概述

企业的存在是通过所创造的产品或服务不断地传达一种影响力，客户需求与企业供给 之间的关系建造是影响力的基础。“互联网+营销”使消费者参与新产品的创造与传播，以 及引导消费者的流变速度和发展速度，进而产生对客户价值的动态认识。在本章所介绍的 推荐模型就是最常用的客户关系管理模型之一。

**16.1** **推荐技术概述**

推荐系统通过建立用户与信息产品之间的二元关系，利用已存有的选择过程或相似性 关系来挖掘每个用户感兴趣的对象，进而进行个性化推荐，其本质就是信息过滤。推荐系 统不仅在社会经济中具有重要的应用价值，而且也是一个重要的科学研究问题。



第16章大数据推荐技术

**16.1.1** **推荐系统的产生与发展**

**1.** **推荐技术的产生**

通过搜索引擎可以解决信息筛选问题，但搜索引擎需要用户主动提供关键词来对海量 信息进行筛选。当用户无法准确描述自己的需求时，严重影响搜索引擎的筛选效果。但是， 将用户的需求转化成关键词的过程并不是一个容易的过程。在这种背景下，推荐技术出现 了，依据推荐技术而建立的推荐系统可以解决上述问题。利用推荐技术， 一方面，帮助用户 发现对自己有价值的信息；另一方面，能够将信息展现在目标人群中。

**2.** **推荐系统的发展过程**

从20世纪90年代推荐系统出现开始，推荐系统发展经历了三个发展阶段。第一阶段 大约从十五年前开始，以提升电商销售为核心，典型的应用是亚马逊的推荐系统。第二阶 段伴随着 Web2.0 的兴起，以围绕用户兴趣推荐为核心，典型的应用是Netfix 的推荐系统， 以及众多 Web 2.0 网站的个性化推荐服务。目前已进入第三个阶段，其特征是以社会网络 和云计算为基础，实现平台化的个性化推荐服务。

不同阶段的个性化推荐服务有着不同的特点、需求和目标，在这里，将比较主流的推 荐算法在不同种类的推荐条目和不同人群上的表现，并以此为基础设计下一代高质量的个 性化推荐系统的基本能力和要求。人人网是中国社交网络的代表，它组织了一个超过2亿 注册用户及月登录用户超过1亿的巨型社交图谱，人人网中的推荐系统是基于大数据的应 用。社会化推荐系统产生的背景就是信息爆炸和信息过载，给用户推荐符合其喜好的信息， 如好友推荐、商品推荐、日志推荐、视频推荐、App 推荐、广告推荐等。

社会化推荐系统有三个主要特征，信息特征、社交特征和流动性。信息特征即多样化、 网状化、异构化实体；社交特性中社交关系是一种特殊的信息；流动性是指信息在社交网 络中动态流动，加速增长，充满噪声。

**16.1.2** **推荐系统的概念**

在互联网的环境下，消费者需要快速方便地找到自己感兴趣的内容，生产者也想将自 己的信息内容推荐给最适合的目标用户，而充当生产者和消费者的中介功能的是推荐系统。 如图16-1所示为推荐系统的概念性示意图。

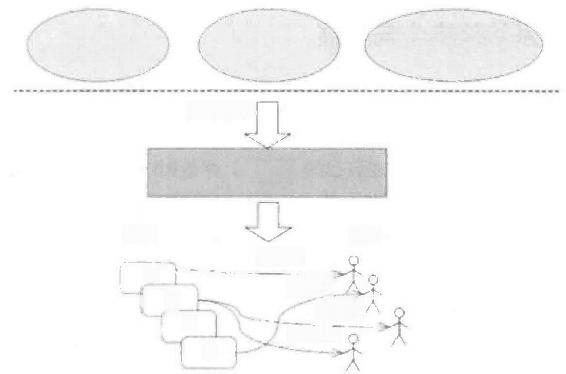
从图16-1 中可以看出，推荐系统的任务就是联系用户和信息，根据物品信息、用户信 息和用户对物品的偏好进行推荐。 一方面帮助用户发现对自己有价值的信息，另一方面将 物品信息能够展现在对他感兴趣的用户中，进而实现产品信息提供商与用户的双赢。

推荐的例子在生活中无处不在，例如买油条的时候，老板就经常问要不要来杯豆浆， 这就是一种简单的推荐。随着互联网的发展，把线下的这种模式搬到线上成了大势所趋， 大大扩展了推荐系统的应用。例如亚马逊的商品推荐系统，Facebook 的好友推荐系统，Digg 的文章推荐系统，豆瓣的豆瓣猜系统和 Gmail 的广告推荐系统等，都是得以广泛应用的推 荐系统。





大数据基础与应用



用户对物品的偏好 (评分，查看，购买…)

数据来源

推荐引擎

用户

推荐给

A

推荐给

推荐给

推荐给

D

物品信息

(关键字，基因…)

用户信息

(性别，年龄…)

物品

B

C

图16-1 推荐系统的概念性示意图

**16.1.3** **推荐系统架构**

推荐系统架构如图16-2所示，主要由三个模块组成。

用户特征

推荐结果 排序

相关物品

检索

提取

图16-2 推荐系统架构

**1.** **用** **户** **特** **征** **提** **取** **模** **块**

用户特征提取模块通过获取用户的历史行为以及其他相关信息生成用户特征，这些特 征可以全面描述用户特征。

**2.** **相** **关** **物** **品** **检** **索** **模** **块**

相关物品检索模块接受用户特征，快速找到用户可能感兴趣的候选物品。这一模块通 过检索的方法实现，通过特征—物品的倒序索引，快速找到与用户相关的物品。

**3.** **推** **荐** **结** **果** **排** **序** **模** **块**

推荐结果排序模块应用机器学习算法，通过优化某一指标得到一个模型，例如可以利 用用户特征和候选物品的特征，计算出用户对物品的预测点击率，然后将候选物品按点击 率排序。预估点击率只是排序参考的一个指标，在很多系统中，排序时还需要综合考虑结 果的多样性和新颖性 。

**16.1.4** **推荐系统类型**

推荐系统建立在拥有大量有效数据基础之上，处理的方法有多种，简述如下。





**第16章** **大数据推荐技术**

**1.** **基于用户行为数据推荐**

互联网上的用户行为众多，从简单的网页浏览到复杂的评价等，蕴含了大量的用户反 馈信息，通过对这些行为的分析，就可以获得用户的兴趣喜好。也可以根据用户对物品的 行为，找出兴趣爱好相似的一些用户，将其中一个用户喜欢的东西推荐给另一个用户。例 如，张老师喜欢看的书有 A 、B 、C 、D, 王老师喜欢看的书有 A 、B 、C 、E, 通过这些数 据就可以判断张老师和王老师对书的选择相似，具有相似的用户行为，于是可给张老师推 荐 E 这本书，同时给王老师推荐 D 这本书。

**2.** **基于用户标签数据推荐**

很多网站在处理物品条目时，将通过用户标注的标签来进行推荐，例如博客的标签云， 豆瓣书影音的标签。这些标签本身就是用户对物品的一种聚类，以此作为推荐的依据。根 据标签推荐物品的基本思想就是找到用户常用的一些标签，然后找到具有这些标签的热门 物品，将其推荐给用户。

如果同时喜欢两个物品的人比较多，就可以认为这两个物品相似，最后只要给用户推 荐与他原有喜好相类似的物品。例如，喜欢看《大数据基础与应用》的人大都喜欢看《Hadoop 基础与应用》和 NoSQL, 如果有人刚看完《大数据基础与应用》,就可以立刻给他推荐 《Hadoop 基础与应用》和 NoSQL。

至于什么时候用基于用户，什么时候用基于物品，这都要根据情况而定。 一般来说， 基于用户更接近于社会化推荐，适用于用户少、物品多、时效性较强的场合；而基于物品 则更接近个性化推荐，适用于用户多、物品少的场合，同时基于物品推荐还可以给出推荐 理由。

用户打标签对于推荐标签相当重要， 一方面能方便用户输入标签，另一方面能提高标 签质量，减少冗余。典型的思想就是将当前物品上最热门的标签和用户自己最常用的标签 综合在一起推荐给用户。在用户标记物品的时候，可以给用户推荐的标签分为“我的标签” 和“常用标签”两类。

基于标签的推荐的优点是， 一方面可以给用户提供比较准确的推荐理由；另一方面标 签云的形式也提高了推荐的多样性，给用户一定的自主选择。可以将标签看作一种物品的 内容数据，例如书的作者、出版社、类型；音乐的国别、风格、作者等，基于这些信息的 推荐可以弥补基于用户行为推荐的不足。

**3.** **基于上下文信息推荐**

上下文是指用户所处的时间、地点、心情等因素，这些因素对于推荐至关重要。例如， 以时间为例，在很多新闻类网站中，时效性异常重要，如果推荐一篇一年前的新闻给用户， 估计会出现笑话。在这种推荐中就需要加入时间衰减因子，对于越久之前的物品，就赋予 越小的权重。同样的思想也可以用在基于用户行为的推荐中，这里有很多可以优化的地方。 对于基于物品的推荐来说，同一用户在间隔很短的时间内喜欢的不同物品可以给予更高的 相似度，而在找相似物品时也可以着重考虑用户最近喜欢的物品。对于基于用户的推荐， 如果两个用户同时喜欢了相同的物品，那么可以给予这两个用户更高的相似度，而在推荐



**大数据基础与应用**

物品时，也可着重推荐选择相近的用户最近喜欢的物品。可以给相似度和用户的行为赋予 一定权重，时间间隔越久，则权重越低，经过这种改进的协同过滤算法往往能得到用户更 满意的结果。

**4.** **基于社交网络数据推荐**

社交网络是产生大数据的重要源泉。实验证明，由于信任的作用，来自朋友的推荐往 往能获取更多的推荐，因此，可以利用信息给用户推荐朋友喜欢的商品。此种推荐类似于 基于用户的推荐，只是寻找用户之间的关系时，除了兴趣相似度以外还得考虑熟悉度，如 共同好友个数，这样一来，朋友们喜欢的物品很可能获得推荐。

在社交网络的内部也应用了许多推荐方法。其中最重要的是朋友推荐，可依据的数据 有很多，例如人口统计学属性，包括共同兴趣、朋友关系(共同朋友数量， N 度人脉)等。 另外还有信息流推荐，其基本思想就是：如果一个会话对所熟悉的朋友最近产生过重要的 行为，它在信息流的排序中就将有比较高的权重。另外，基于社交网络兴趣图谱和社会图 谱的精准广告投放也是推荐系统的关键应用，它决定了社交网站的能力。

**16.1.5** **推荐系统的评判标准**

可以通过如下几个标准来判定推荐系统。

**1.** **用户满意度**

用户满意度描述了用户对推荐结果的满意程度，这是推荐系统最重要的指标。 一般通 过对用户进行问卷或者监测用户线上的行为数据获得。

**2.** **预测准确度**

描述推荐系统预测用户行为的能力。 一般通过离线数据集上算法给出的推荐列表和用 户行为的重合率来计算。重合率越大，则准确率越高。

**3.** **覆盖率**

描述推荐系统对物品长尾的发掘能力。 一般通过所有推荐物品占总物品的比例和所有 物品被推荐的概率分布来计算。比例越大，概率分布越均匀，则覆盖率越大。

**4.** **多样性**

描述推荐系统中推荐结果能否覆盖用户不同的兴趣领域。 一般通过推荐列表中物品两 两之间不相似性来计算，物品之间越不相似，则多样性越好。

**5.** **新颖性**

如果用户没有听说过推荐列表中的大部分物品，则说明该推荐系统的新颖性较好。可 以通过推荐结果的平均流行度和对用户进行问卷来验证新颖性。





第 1 6 章 大数据推荐技术

**6.** **惊喜度**

如果推荐结果和用户的历史兴趣不相似，但用户很满意，则可以说这是一个让用户惊 喜的推荐。可以定性地通过推荐结果与用户历史兴趣的相似度和用户满意度来衡量。

**7.** **冷启动**

一个以推荐功能为主体的网站，在其创立之初的最主要问题就是冷启动问题。在刚创 建的网站，无用户的积累，也无物品的积累，推荐就无规律可循，这就是推荐系统的冷启 动问题。冷启动问题是推荐系统的一个重要问题，冷启动问题可细分为下述三种情况。

● 如何给新用户做个性化推荐?

● 如何将新物品推荐给用户?

● 在数据稀少的情况下，新网站如何完成个性化推荐，即推荐内容从何而来?在没 有行为时如何做个性化推荐?

对于新用户，首先可以根据其注册信息进行粗粒度的推荐，如年龄、性别、爱好等。 另外也可以在新用户注册后为其提供一些内容，使他们反馈对这些内容的兴趣，再根据这 些数据来进行推荐。这些内容需要同时满足热门和多样的要求。而对于新物品的推荐，可 以通过语义分析对物品抽取关键词并赋予权重，这种内容特征类似一个向量，通过向量之 间的余弦相似度便可得出物品之间的相似度，从而进行推荐。这种内容过滤算法在物品(内 容)更新较快的服务中得到大量应用。

在初建网站，数据量不大的情况下，是通过人工的方法来建立初始的推荐系统。较简 单的是人工编辑热门榜单，较高级的是人工分类多维度标注。有了这些初始数据，就可以 方便地进行推荐了。除此之外，也可利用社交网络平台的已有大量数据，尤其是那些依托 于其他SNS 系统的服务。

总之， 一个好的推荐系统就是在推荐准确的基础上，给所有用户推荐的物品尽量广泛， 给单个用户推荐的物品尽量覆盖多个类别，同时不要给用户推荐太多。

**16.2** **推荐算法与推荐模式**

**16.2.1** **推荐算法**

虽然推荐算法不能解决全部问题，但推荐算法可以使基于推荐算法书写的推荐软件变 得更加智能化和人性化。

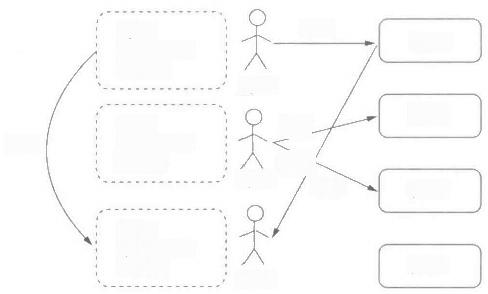
**1.** **基于人口统计学的推荐算法**

基于人口统计学的推荐算法是最为简单的一种推荐算法，如图16-3所示，它只是简单 地根据系统用户的基本信息发现用户的相关程度，然后将相似用户喜爱的其他物品推荐给 当前用户。





大数据基础与应用



用户A

年龄：25-30

性别：女

用户A

用户B

年龄：30-35

性别：男

物品C

用户C

年龄：25-30

性别：女

用户C

喜欢

喜欢

物品A

物品D

物品B

用户B'

推荐

喜欢

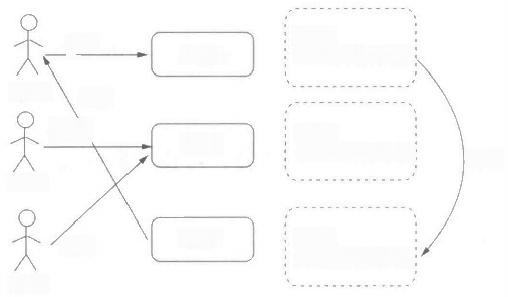
相似

图16-3 基于人口统计学的推荐算法

系统首先根据用户的年龄、性别和兴趣等用户的属性信息建模。根据这些特征计算用 户间的相似度。比如系统通过计算发现用户A 和 C 较相似，就将A 喜欢的物品推荐给C。 基于人口统计学的推荐算法的缺点是算法比较粗糙，只适用简单的推荐，效果不佳。基于 人口统计学的推荐算法的优点是不需要历史数据，无冷启动问题，不依赖于物品的属性， 因此其他领域的问题都可无缝接入。

**2.** **基于内容的推荐算法**

基于内容的推荐算法如图16-4所示，与人口统计学的算法相类似，只不过这次的中心 转到了物品本身。使用物品本身的相似度而不是用户的相似度进行推荐，即物品本身的相 似度是通过计算它们属性之间的相似度来确定。



喜欢

推荐

喜欢

类型：恐怖，惊悚：

用户B

喜欢

用户C

电影A

类型：爱情，浪漫

电影C

类型：爱情，浪漫

电影C

电影A

电影B

电影B

用户A

相似

图16-4 基于内容的推荐算法

图16-5中是以电影内容为例，系统首先对物品的属性进行建模，用类型作为属性。在 实际应用中，只根据类型显然过于简单，还需要考虑演员、导演等更多信息。通过相似度 计算，发现电影A 和 C 相似度较高，因为它们都属于爱情类。系统还会发现用户A 喜欢电 影 A, 由此得出结论，用户A 很可能对电影C 也感兴趣。于是将电影 C 推荐给A。基于内 容的推荐算法的优点是对用户兴趣可以很好地建模，并通过对物品属性维度的增加，获得





第16章 大数据推荐技术

更好的推荐精度。该算法存在下述不足

● 物品的属性有限，很难有效地得到更多数据。

● 物品相似度的衡量标准只考虑到了物品本身，有一定的片面性。

● 需要用户的物品的历史数据，有冷启动的问题等。

**3.** **协同过滤推荐算法**

(1)协同过滤推荐算法的内容

作为个性化推荐系统核心的协同过滤算法，是高德博格(Goldberg) 等人在1992年最 早提出。在一个新闻组中，根据用户下载的新闻，计算新闻之间在选择上的相似程度，并 利用这种相似程度为他们进一步推荐相关的新闻。这也是个性化推荐系统的雏形。

协同过滤推荐算法是推荐算法中最经典最常用的算法，协同过滤推荐与传统的基于内 容过滤直接分析内容进行推荐不同，协同过滤分析用户兴趣，在用户群中找到指定用户的 相似(兴趣)用户，综合这些相似用户对某一信息的评价，形成系统对该指定用户对此信 息的兴趣程度预测。协同过滤推荐是一项很受欢迎的技术。

协同过滤算法分为两种，分别为基于用户和基于物品的协同过滤算法。

① 基于用户协同过滤算法。基于用户协同过滤算法就是根据用户对物品的行为，找出 兴趣爱好相似的一些用户，将其中一个用户喜欢的东西推荐给另一个用户。协同过滤算法 集中关注用户和物品之间的关系，物品之间的相似度是通过同时对它们进行评级的用户所 评级的所有项的评级结果之间的相似度来确定。

② 基于物品协同过滤算法。基于物品协同过滤算法的运行过程是首先找出相似的物 品，相似物品的判定方法是如果同时喜欢两个物品的用户比较多，就可以认为这两个物品 相似，在推荐时，就将与原有喜好相似的物品推荐给用户。

协同过滤算法也有不少缺点，最明显的一个就是热门物品的干扰。举个例子，协同过 滤算法经常会导致两个不同领域的最热门物品之间具有较高的相似度，这样很可能会给喜 欢《算法导论》的同学推荐《哈利·波特》,显然，这是毫无根据的推荐。要避免这种情况 就得从物品的内容数据入手，基于内容过滤算法就是其中一种。除了协同过滤算法，还有 隐语义模型应用得也比较多，它基于用户行为对物品进行自动聚类，从而将物品按照多个 维度、多个粒度分门别类，然后根据用户喜欢的物品类别进行推荐。这种基于机器学习的 方法在很多指标上优于协同过滤，但性能上不太理想， 一般可以先通过其他算法得出推荐 列表，再由隐语义模型进行优化。

(2)协同过滤算法的优点

● 能够过滤难以自动基于内容分析的信息。

● 能够基于一些复杂的，难以表达的概念(信息质量、品位)进行过滤。

● 推荐具有开放性和新颖性。可以共用他人的经验，很好地支持用户发现潜在的兴 趣爱好。

● 不需要对物品或者用户进行严格的建模，而且不要求物品的描述是机器可理解的， 所以这种方法与领域无关。

协同过滤已成功地应用在商务上。Amazon、CDNow、MovieFinder 等都采用了协同过





**大数据基础与应用**

滤的技术来提高服务质量。

(3)协同过滤算法的缺点

● 用户对商品的评价非常稀疏，这样基于用户的评价所得到的用户间的相似性可能 不准确，即存在稀疏性问题。

● 随着用户和商品的增多，系统的性能会越来越低。

● 如果从来没有用户对某一商品加以评价，则这个商品就不可能被推荐。

● 方法的核心是基于历史数据，所以对新物品和新用户都存在冷启动的问题。

● 推荐的效果依赖于用户历史偏好数据的多少和准确性。

● 在大部分的实现中，用户历史偏好是用稀疏矩阵进行存储的，而稀疏矩阵上的计 算有些明显的问题，包括可能少部分人的错误偏好会对推荐的准确度有很大的影 响等。

● 对于一些特殊品位的用户不能给予很好的推荐。

● 由于以历史数据为基础，抓取和建模用户的偏好后，很难修改或者根据用户的使 用演变，从而导致这个方法不够灵活。

现在的电子商务推荐系统都采用了几种技术相结合的混合推荐技术。

**4.** **算法比较**

(1)基于用户的协同过滤与基于人口统计学的推荐的比较

基于用户的协同过滤推荐机制和基于人口统计学的推荐机制都是计算用户的相似度， 并基于邻居用户群计算推荐，但不同的是如何计算用户的相似度，基于人口统计学的机制 只考虑用户本身的特征，而基于用户的协同过滤机制可是在用户的历史偏好的数据上计算 用户的相似度，它的基本假设是喜欢类似物品的用户可能有相同或者相似的偏好。

(2)基于物品的协同过滤与基于内容的推荐的比较

基于项目的协同过滤推荐和基于内容的推荐其实都是基于物品相似度预测推荐，只是 相似度计算的方法不一样，前者是从用户历史的偏好推断，而后者是基于物品本身的属性 特征信息。

**5.** **混合推荐算法**

以上介绍的方法是最常见的几种推荐方法。可以看出，每个方法都不完美。因此在实 际应用中，混合推荐算法更受欢迎，它能各取所长。在实际应用时，可以考虑一下当前情 况下更适合什么算法，进而提高系统的效率。

**16.2.2** **推荐模式**

**1.** **推荐模式类型**

根据出发点的不同，推荐系统可分为基于协同过滤和基于内容的两种推荐系统。从实 际应用的角度，这两种推荐系统所使用的模式分别为黑盒推荐模式和白盒推荐模式。





第 1 6 章 大数据推荐技术

(1)黑盒推荐模式

黑盒推荐不需要考虑推荐的具体内容，而是利用机器学习、数据挖掘等统计方式和人 工智能的方法对数据进行分析，建立相关模型和优化目标，在一定约束条件下求得最优解 或局部最优解，以此作为向用户推荐的内容。

(2)白盒推荐模式

白盒推荐深入到被推荐的条目内容之中，依据对条目的先验知识和对用户的理解进行 相关匹配的推荐，推荐的过程中也会用到机器学习和数据挖掘的算法，但先验知识的来源 往往是专家领域的知识。

(3)黑盒推荐模式的弊端

比较黑盒模式与白盒模式的优劣可以看出，在实际应用中，通常融合两种模式的优势 建立混合模型以求得更好的推荐效果。考虑到可扩展性及系统建立的成本，大型的个性化 推荐系统大多数以黑盒推荐模式为基础。黑盒推荐模式的核心是机器学习和数据挖掘，具 有坚实的数学基础和明确的优化指标与方法，这是对推荐质量的基本保证。构建系统既不 需要领域内的知识，也不需要过多的人工干预，同时模型的鲁棒性比较好，对用户和条目 的增长不需要付出更多额外的努力。这些优势使得黑盒推荐模型获得了非常广泛的应用， 但随着前面提到的互联网领域的两个根本性变化的影响，黑盒推荐模式的弊端也越来越多 地显现出来，如图16-5所示。



用户收藏

图16-5 黑盒推荐模式的弊端

以协同过滤为基础的推荐算法的主要弊端是冷启动问题，即数据稀少的情况下难以获 得高质量的推荐。这是黑盒推荐模式在系统启动时面临的最大问题。事实上，即便有了启 动数据，在数据稀疏的情况下，黑盒推荐模式偶尔会将噪声放大，给出低质量的推荐。例 如在某著名电子商务网站上，每年母亲节时都有鲜花促销，十几岁的青少年在为母亲购买 鲜花之余，顺便会为自己买上几部功夫片。系统中的鲜花和功夫片便建立了相似关联，在 有人购买鲜花时系统便会推荐功夫片，反之亦然。 这种推荐自然会引起很多用户的不适， 虽然从统计意义的优化指标来看系统达到了最优，但对于某些主观性的因素，例如用户对 系统推荐的信任和心理感受，却存在着负面的影响。这些影响可能很难用具体的指标和数 字来衡量，但是它们对一个产品的成败来说，往往是长期的、根本性的，甚至是决定性的。

**2.** **黑盒推荐模式的三个阶段**

图16-5描述了黑盒推荐模式的推荐质量和用户收藏之间的关系，可以比较清晰地将其 划分为如下三个阶段。





**大数据基础与应用**

(1)第一阶段

用户收藏较少，系统在宏观和微观层面都面临冷启动问题，此时的推荐质量低于用户 的预期。这是用户在使用个性化推荐系统时的第一个门槛，将把相当一部分用户阻挡在产 品外。

(2)第二阶段

随着用户收藏的增长，系统对用户兴趣的建模更加准确，推荐质量也获得明显提升， 这段时间推荐系统和用户之间的紧密结合，此时的推荐往往会给用户带来满意的结果，能 够有效地帮助用户发掘他们未知但感兴趣的领域。

(3)第三阶段

这个阶段也是宏观和微观的悖论所产生的阶段。此时，虽然从各项指标来看推荐质量 还在继续改善，但用户所切身感受到的却是推荐系统能给他们提供的帮助越来越少，推荐 内容趋于平庸。

针对黑盒推荐模式的弊端，提出了很多克服的方案。例如结合基于内容的白盒推荐模 式方法来解决冷启动问题从而提高推荐的质量、通过变换的方法利用其他领域的结果和训 练的模型获得启动数据，以及利用矩阵分解的方法改善数据稀疏性的问题。这些方法都收 到了不错的效果，值得实践领域的从业者重视。

**16.2.3** **下一代推荐系统**

**1.** **推荐系统面临的问题**

在21世纪，互联网领域两个根本性的变化使得推荐系统需要面临新的问题。 (1)Web 2.0 的兴起

第一个变化是随着 Web 2.0 的兴起，个人用户逐渐成为网站的中心。用户喜欢在网络 上建立和分享他们的社会关系和兴趣爱好，进而展示个性。网站的创立者也更重视对用户 的基础数据的收集和分析，从而全方位地满足用户的需求。与传统的以信息为中心的 Web 1.0的组织方式相比， Web 2.0强调以用户为中心，对个性化推荐系统的发展起到了极大的 促进作用。因此，几乎每个具有一定规模的Web 2.0网站都会建立用户模型，甚至还涌现出 不少以个性化推荐为核心的服务。这使得个性化推荐系统的研究和实践获得了显著的进展。

(2)互联网深入真实生活

第二个变化是互联网越来越深入真实生活，人们越来越习惯在网络上使用真实身份， 维持真实的社会关系，分享真实的生活轨迹。互联网不再是虚拟世界的代表，而是真实世 界的一部分。这使得网站的创立者能够更准确地掌握用户的各种信息。而用户对个人隐私 的关注也使得工程师在使用这些信息时不得不更加慎重。

对个性化推荐系统来说，这些变化提供更多的真实信息有利于提高推荐精度；而同时 人们对隐私和信息流动方向的关注为推荐系统如何使用这些信息设置了更高的门槛。个性 化推荐系统不但要考虑推荐的精度，还要考虑在社会网络中信息的来源和用户对这些信息 微妙的情感因素，而这些因素在算法中往往是难以建模和衡量的。





第16章 大数据推荐技术

**2.** **推荐系统主要考虑的问题**

● 需要考虑在保证推荐准确性的同时增加推荐的多样性，使特殊喜好的用户得到 推荐。

● 需要考虑用户对系统的实时反馈，更好地利用反馈数据强化推荐算法。

● 考虑时间、空间、任务等因素，来更好地完成推荐。

黑盒推荐模式面临一个根本问题是推荐系统过分关注短期的优化目标或产品目标，忽 视了产品的用户价值和增长的原动力。其表现就是没有把用户在领域内的动态成长模型纳 入到算法框架之中。

**3.** **推荐系统的层次结构**

好的个性化推荐算法，应该是始于用户、终于用户的，并且是一个健康的、动态的、 具有自我生长和自我调节能力的系统。用户是系统的一部分，与系统共同促进和成长。依 据这样的思想，可以把推荐系统分成猜、预测和推荐三个层次，如图16-6所示。



猜 预测 推荐

图16-6 下一代个性化推荐系统的层次结构

(1)猜层次

在猜层次中，猜就是根据部分静态的数据，想办法恢复出全部数据，使得恢复误差做 到最小。这部分工作可以认为是个数学问题，能够形式化并给出严格的定义，这也是目前 研究得最透彻、解决得最好的一部分。这部分工作也是建立个性化推荐系统的第一层次， 其意义是从数据形成信息。

(2)预测层次

预测是指预测系统下一步的变化。用上一个层次的方法也可以解决这一层次的问题， 但其中有一个重要的区别，即在时间维度上预测要解决的问题在系统之外，而猜要解决的 在系统之内。这使得预测比猜在优化指标的设置上要复杂一些。同时，因为时间方向的不 同，所以在系统中要考虑人的心理和社会环境因素所起的作用。这两个因素的作用使预测 的系统比上一个层次的复杂度高了很多，原来的方法可行，但未必有效。解决预测的问题， 仅从给定的数据出发不够，需要整合其他来源的数据，可能需要先验知识，甚至需要用户 的认知和心理模型。

(3)推荐层次

在推荐层次，推荐是指真正意义的个性化推荐。这个系统应该具备所推荐领域的知识， 同时还要认识它的每一个用户。这个系统能够自我启动，同时能够生长和进化。这个系统 具有记忆能力，同时还具有学习能力。这个系统没有严格的定义，同时它的基本技术架构 可能也不会仅局限于目前流行的个性化推荐系统的各种框架、技术和算法之中。随着计算 能力的飞速发展和数据的极大丰富，将出现全新的、突破性的进展。





**大数据基础与应用**

**16.3** **大数据推荐技术的挑战**

大数据的出现，虽然为推荐技术带来了机遇，但又带来了新的问题，下面仅介绍10个 典型问题。

**16.3.1** **数据稀疏性问题**

虽然用户和商品(也包括其他物品，如音乐、网页、文献等)数目巨大，但两个用户 之间选择的相重叠内容却非常少。如果用用户和商品之间已有的选择关系占所有可能存在 的选择关系的比例来表示稀疏度，可以用稀疏度来衡量系统的稀疏性，稀疏度的数值越小， 则稀疏性越强。数据集的稀疏度是1.2%～4.5%,这已经是非常密集的数据了。淘宝上号称 有8亿商品，平均一个用户能够浏览800件商品已不可能，所以稀疏度应该在百万分之一 或百万分之一以下的量级。数据非常稀疏，影响了绝大部分基于关联分析的算法(如协同 过滤)的效果。为了解决这个问题，提出了很多方法。例如可以通过扩散的算法，从原来 的一阶关联(两个用户有多少相似打分或者共同购买的商品)到二阶甚至更高阶的关联(假 设关联性或者相似性本身是可以传播的),也可以添加一些默认项，从而提高相似性的分辨 率。 一般说，数据规模越大，则越稀疏，因此处理稀疏数据的算法十分重要。

**16.3.2** **大数据冷启动**

由于新用户能够的行为信息稀少，所以很难给出精确的推荐。反过来，新商品由于被 选择次数很少，也难以找到合适的办法推荐给用户。解决这种问题的方法是利用文本信息 进行辅助推荐，或者通过注册以及询问得知一些用户的属性信息，如年龄、居住城市、受 教育程度、性别、职业等。标签系统的应用提供了解决冷启动问题的一种可能方案，因为 标签既可以看作是商品内容的抽取，同时也反映了用户的个性化喜好。对于同一商品，用 户兴趣点可能相同，打上的标签也不同。当然，利用标签也只能是提高具有少量行为的用 户的推荐准确性，对于纯粹的冷启动用户无帮助，因为这些用户还没有打过任何标签。新 用户更容易选择特别流行的商品，这就说明了应用热销榜也能获得不错的结果。冷启动问 题也可以通过多维数据的交叉推荐来解决，其精确度和多样性又远优于热销榜。显然，与 大数据技术密切相关，这是因为物品数量和用户数量剧增，导致冷启动问题增大。

**16.3.3** **多样性与精确性的两难命题**

如果要给用户推荐所喜欢的商品，最保险的方式就是推荐特别流行或者得分特别高的 商品，因为这些商品有更大的可能性被推荐成功，但是，这样的推荐产生的用户体验并不 一定好，因为用户很可能已经知道这些热销流行的产品，所以得到的信息量很少，并且用 户不会认同这是一种个性化的推荐。事实上，盲目崇拜精确性指标可能会伤害推荐系统，





**第16章** **大数据推荐技术**

因为这样可能会导致用户得到一些信息量为0的精准推荐，并且视野变得越来越狭窄。让 用户视野变得狭窄也是协同过滤算法存在的一个主要的缺陷。与此同时，应用个性化推荐 技术也希望推荐中有更多的品类出现，从而激发用户新的购物需求。但是，推荐多样的商 品和新颖的商品与推荐的精确性之间存在矛盾，因为前者风险很大， 一个没什么人看过或 者打分较低的物品被推荐，很可能不被用户喜欢，从而效果更差。很多时候，这是一个两 难的问题，即只能通过牺牲多样性来提高精确性，或者牺牲精确性来提高多样性。 一种可 行之策是直接对推荐列表进行处理，从而提升其多样性，这种方法固然在应用上有效，但 是没有任何理论的基础可言，只能算一种野蛮而实用的方法。通过精巧混合精确性高和多 样性好的两种算法，可以同时提高算法的多样性和精确性，不需要牺牲任何一方。

**16.3.4** **大数据处理与增量计算问题**

尽管数据很稀疏，但都拥有海量用户和商品，因此，如何快速高效处理这些数据成为 迫在眉睫的问题，而算法时间和空间的复杂性，尤其是算法时间的复杂性，获得了空前重 视。 一个高效的算法，主要体现在复杂性很低，或者高度并行化，要么两者兼备。局部扩 散算法在这两个方面都具有明显优势，另外一条解决方法是增量算法，即当产生新用户、 新商品以及新的连接关系时，算法的结果不需要在整个数据集上重新进行计算，而只需要 考虑所增加节点和连边局部的信息，对原有的结果进行微扰，快速得到新结果。这种增量 算法随着加入的信息量的增多，其误差会积累变大，每过一段时间需要利用全局数据重新 进行计算。 一个特别困难的挑战是如何设计一种算法，能够保证其误差不会累积，也就是 说其结果与利用全部数据重新计算的结果之间的差异不会单调上升，把这种算法叫作自适 应算法，它比增量算法设计要求和难度更高。增量算法已经在业界有了应用，例如百分点 科技推荐引擎中的若干算法都采用了增量技术，使得用户每次新浏览收藏或者购买商品后 其推荐列表立刻得到更新。当然，该引擎也只是部分算法实现了增量技术，更没有达到所 有算法都能够自适应学习的程度。

**16.3.5** **推荐系统的脆弱性问题**

受推荐系统在电子商务领域重大的经济利益的驱动， 一些心怀不轨的用户通过提供一 些虚假恶意的行为，故意增加或者压制某些商品被推荐的可能性。因此， 一个算法能否在 一定程度上保持对恶意攻击的鲁棒性，成为需要认真考虑的一个特征。以最简单的关联规 则挖掘算法为例，通过分析对比真实用户和疑似恶意用户之间打分行为模式的差异，提前 对恶意行为进行判断，从而阻止其进入系统或赋予疑似恶意用户比较低的影响力。总体来 说，这方面的研究相对较少，系统性的分析还很缺乏，反而是攻击策略层出不穷。

**16.3.6** **推荐系统效果评估**

推荐系统的评估指标可以分为四大类，分别是准确度、多样性、新颖性和覆盖率，每 一类下辖很多不同的指标。



**大数据基础与应用**

准确度指标又可以分为四大类，分别是预测评分准确度、预测评分关联、分类准确度、 排序准确度四类。以分类准确度为例，又包括准确率、召回率、准确率提高率、召回率提 高率、F1 指标和 AUC 值。几乎所有的推荐系统指标都是基于数据本身的指标，可以认为 是第一个层次。实际上，在实际应用时，更为重要的是另外两个层次的评价，第二个层次 是商业应用上的关键表现指标，例如受推荐影响的转化率、购买率、客单价、购买品类数 等，第三个层次是用户真实的体验。绝大部分研究只针对第一个层次的评价指标，而业界 真正感兴趣的是第二个层次的评价(如到底是哪个指标或者哪些指标组合的结果能够提高 用户购买的客单价),而第三个层次最难，没人能知道，只能通过第二个层次来估计。如何 建立第一层面和第二层面指标之间的关系，就成为关键，如果解决了，理论和应用之间的 屏障就通了一大半。

**16.3.7** **用户行为模式的挖掘和利用**

深入挖掘用户的行为模式可以提高推荐的效果或在更复杂的场景下进行推荐。例如， 新用户和老用户具有不一样的选择模型，新用户倾向于选择热门的商品，而老用户对于小 众商品关注更多，新用户所选择的商品相似度更高，老用户所选择的商品多样性较高。有 些混合算法可以通过一个单参数调节推荐结果的多样性和热门程度，在这种情况下就可以 考虑给不同用户赋予不同参数(从算法结果的个性化到算法本身的个性化),甚至允许用户 自己移动一个滑钮调节这个参数，当用户想看热门商品的时候，算法提供热门推荐。当用 户想找点很酷的产品时，算法也可以提供冷门推荐。用户行为的时空统计特性也可以用于 提高推荐或者设计针对特定场景的应用。用户的选择可能同时蕴含了长期的兴趣和短期的 兴趣，通过将这两种效应分离出来，可以明显提高推荐的精确度。事实上，简单假设用户 兴趣随时间按照指数递减，也能够得到改进的推荐效果。利用手机上网现在已经越来越普 及，与此同时，嵌入GPS 的手机运用，基于位置的服务成为一个受到广泛关注的问题。基 于位置信息的推荐可能会成为个性化推荐的一个研究热点和重要的应用场景，而这个问题 的解决需要能够对用户的移动模式有深入理解，包括预测用户的移动轨迹和判断用户在当 前位置是否有可能进行餐饮购物活动等，同时还要有定量的办法去定义用户之间以及地点 之间的相似性。另外，不同用户打分的模式也很不一样，用户针对不同商品的行为模式也 不一样，这些都可以用来提高推荐的效果。

**16.3.8** **用户界面与用户体验**

这个问题是实际应用的问题，推荐结果的可解释性，对于用户体验有至关重要的影响， 用户希望知道这个推荐是怎么来的。在这个意义上，协同过滤有明显的优势，例如亚马逊 基于商品的协同过滤在发送推荐的电子邮件时会告诉用户之所以向其推荐某书，是因为用 户以前购买过某些书。相对地，矩阵分解或者集成学习算法很难向用户解释推荐结果的起 源。用户则更喜欢自己朋友的推荐而不是系统的推荐。另外，推荐列表往往含有很多项， 这些推荐项最好能够区分成很多类别，不同类别往往来自不同的推荐方法。如何更好地呈





**第16章** **大数据推荐技术**

现推荐是一个很难建立理论模型和进行量化的问题，对于不同被推荐品而言，用户界面设 计的准则也可能大不相同。

**16.3.9** **多维数据的交叉利用**

网络科学的重要概念是相互作用的网络结构和动力学。网络与网络之间的相互作用可 以分成三类。

**1.** **依存关系**

例如电力网络和 Internet, 如果发生了大规模停电事故，当地的自主系统和路由器也会 受到影响，导致网络局部中断。

**2.** **合作关系**

例如人的一次出行，可以看作航空网络、铁路网络和公路网络的一次合作。

**3.** **交叠关系**

主要针对社会网络，几乎每一个人，都参与了不止一个大型的社会网络中，例如可能 既有新浪微博的账号，又是人人网的注册用户，还是手机用户，那么他已经同时在三个巨 大的社会网络中了。与此同时，此人可能还经常在淘宝、京东、麦包包、库巴网等网站进 行网购，那么他也是一张巨大的用户一商品图中的一员。如果能够把这些网络数据整合起 来，特别是知道每个节点身份的对应关系，不需要知道真实身份，只需要知道不同网络中 存在的一些节点是同一个人，将具有巨大的社会经济价值。

交叠社会关系中的数据挖掘称为多维数据挖掘，是解决系统内部冷启动问题的有效方 法。只要用户在系统外部的其他系统有过活动，就可以利用用户在其他电商的浏览购买历 史为提高在目标电商推荐的精确度。与此同时，可以利用微博和其他社会网络的活动提高 商品推荐的精度，还可以反过来利用商品浏览历史提高微博关注对象推荐的精度。这样就 不会陷入总在一个圈子里面来回推荐的弊端。

**16.3.10** **社会推荐**

研究结果表明，用户更喜欢来自朋友的推荐而不是被系统计算出的推荐。社会影响力 比历史行为的相似性更加重要，通过社会关系的分析，可以大幅度提高推荐的精确度。来 自朋友的社会推荐有两方面的效果： 一方面是增加销售(含下载、阅读等),另一方面是在 销售后提高用户的评价。当然，社会推荐的效果也不完全是正面的，例如朋友推荐对书的 销售增长几乎没有帮助，有时候还会起到负面作用。在社会推荐方面存在的挑战主要可以 分为三类。

● 如何利用社会关系提高推荐的精确度?

● 如何建立更好的机制以促进社会推荐?

● 如何将社会信任关系引入到推荐系统中?





**大数据基础与应用**

社会推荐的效果可能来自类似口碑传播的社会影响力，也可能是因为朋友之间本来就 具有相似的兴趣或者兴趣相投更易成为朋友，对这些不同的潜在因素进行量化区别，也需 要进一步研究。

**16.4** **大数据人才推荐系统**

基于大数据的精准员工推荐模式及解决方案是企业定制化的招聘需求，通过对社会化 媒体及简历数据库中用户关系和文本描述大数据的定向挖掘，帮助人力资源主管通过社交 招聘这一全新模式成功实现精准化、智能化、个性化的员工推荐和筛选，使招聘工作变得 更为简单、高效和有趣。

在招聘工作中，如何从成千上万份简历中挑选出合适的应聘者，是企业人力资源工作 面临的最大挑战之一。基于大数据分析而产生人才招聘系统，不同于传统的人力资源办公 管理系统，它面向的是招聘的业务支撑和决策系统，并结合大数据社交网络数据挖掘和分 析，提出一套同时面向求职者和招聘者的双向扩展匹配算法，既能让人力资源工作人员在 茫茫人海中发现与职位需求高度匹配的专业人才，又能为求职者提供个性化的营销渠道， 找到能够展现个人才华的最佳舞台。人才招聘系统的独特优势还体现在对各种社交网络大 数据的深度分析上，通过建立求职者的性格图谱、兴趣图谱和关系图谱，深入了解求职者 的性格特点、兴趣方向和社交圈子。这些因素在招聘方最终决策和求职者人生职业规划中 都占有重要地位。同时，基于大数据的人才招聘系统在企业内部员工推荐上的优势更为明 显，通过对员工社交关系的延展和判断，系统不仅能精准发现与职位需求匹配的求职者， 还能计算求职者与推荐者的信任关系及参与应聘的意愿度。

目前，越来越多的企业意识到数据在人力资源管理中的重要性，并希望将其运用到 人力资源战略决策中，通过数据分析和量化招聘渠道对人才引进产生影响，并建立有效 的人才数据库。管理并持续积累人力资源方面的战略数据资产，是众多企业当前遇到的 普遍挑战。

据分析，传统的商业智能系统中用以分析人力资源的数据，大都是企业自身信息系统 所产生的标准化和结构化的运营数据，低于企业可利用数据的30%,另外70%的非结构化 和半结构化数据则广泛存在于以社交网络为代表的媒介之中。这也表明企业一旦掌握了基 于社交网络等媒介所产生的定向人才数据，就能够掌握获取优秀人才的机会。

基于社交媒体的广泛应用及可预见的巨大市场需求，行业专家和数据科学家设计和开 发出了人才招聘系统，其自动采集各大社交媒体和专业网站的多源异构海量数据，利用文 本挖掘、网络分析、机器学习等算法，将人才特征和企业招聘需求进行智能匹配，通过云 服务平台，帮助企业的人力资源工作人员找到和推荐合适的候选人，成为支撑企业招聘业 务不可或缺的外脑。

大数据时代的意义不仅在于数据体量之大，更在于大量的数据能够帮助企业进行科 学决策，提升企业效率。通过引入外部数据源以及最先进的算法，改变了传统人力资源 管理中缺乏数据，凭直觉与经验进行决策的不足。这一方式将推动传统的推荐系统的变





第16章 大数据推荐技术

革与创新。

小结

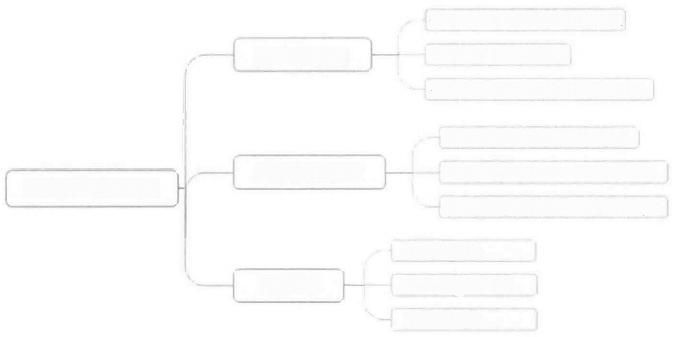
随着 Internet的广泛应用，推荐系统飞速发展，其功能日益强大。本章主要介绍了推荐 技术的基础和主要概念，并较详细地说明了较流行的推荐算法和推荐模式及系统。对于下 一代推荐系统，尤其是基于大数据的推荐系统问题做了进一步的介绍，主要包括推荐系统 的冷启动问题、数据稀疏性问题、多样性与精确性的两难命题等一系列问题。



**第17章**

本章主要内容

**数据科学与数据思维**



数据科学的定义与信息化过程

数据科学的研究内容

数据科学的研究过程与体系框架

大数据分析的是全面的数据

重视数据的复杂性与弱化精确性

数据科学与数据思维

关注数据的相关性而非因果关系

数据科学家

数据处理工程师

大数据思维

大数据研究方式

数据科学概述

数据专家

世界已经从一个依靠模型和假设、数据奇缺的时代，进入到一个模型和假设已逐渐清 晰的大数据的时代。计算机科学是算法与算法变换的科学，数据科学研究范围更为广泛。 数据科学与人工智能和统计学不同，数据科学研究与进展不仅可以推动数学、计算机科学、 统计学、天体信息学、生物信息学、计算社会学等学科的发展，而且能够大力助推产业发展。

**17.1** **数据科学概述**

数据是描述事物的符号记录，可以将其定义为有意义的实体，即它涉及事物的存在形 式。更进一步说，数据是关于事件的一组离散且客观的事实描述，是构成信息和知识的原 始材料。数据可分为模拟数据和数字数据两大类。显然，数据是计算机处理的原始材料， 包括图形、声音、文字、数值、字符和符号等都可以作为计算机的输入数据。

**17.1.1** **数据科学的定义与信息化过程**

**1.** **数据科学的定义**

数据科学是指导数据分析与处理的科学、系统的理论与方法，数据科学通过系统性地 研究数据的组织和使用，可以促进发现、改进关键决策过程。数据科学是研究数据的科学， 研究数据界的理论、方法和技术，研究的对象是数据界中的数据。数据科学将数据作为一个自



**第17章** **数据科学与数据思维**

然体来研究，也就是脱离各个领域的物理世界。物理世界在网络空间中有其数据映象，研究数据 界的规律其实就是网络空间中共同的规律。

数据科学是利用计算机的运算能力与存储能力对数据进行处理，从数据中提取信息，

进而形成知识。数据科学主要有两个内涵： 一个是数据本身，研究数据的各种类型、状态、 属性及变化形式和变化规律；另一个是为自然科学和社会科学研究提供一种新的方法，称 之为科学研究的数据方法，其目的是揭示自然界和人类行为现象和规律。数据科学已经影 响了计算机视觉、信号处理、自然语言识别等计算机分支，且在IT 、金融、医学、交通、 销售等领域得到了广泛使用。

**2.** **产生数据的过程**

产生数据的过程是一个信息化过程，是将现实世界中的事物和现象以数据的形式存储 到存储空间中。这些数据不仅是自然与生命的一种表示形式，而且还记录了人类的行为， 包括工作、生活和社会发展。当前，各类型数据快速、大量地产生并存储在存储空间中， 早期将这种现象称为数据爆炸，现在称为大数据。研究和探索存储空间中数据的规律和现 象，是探索宇宙的规律、探索生命的规律、寻找人类行为的规律、寻找社会发展的规律的 一种重要手段。

**17.1.2** **数据科学的研究内容**

数据科学的研究内容包括基础理论研究、数据技术及其应用研究及数据科学的学科体 系研究。数据科学学科建立，需要对知识结构、课程设置、专业设置等进行学科体系建设， 探讨数据科学、自然科学和社会科学之间的关系，数据科学、计算机科学和信息科学之间 的关系等。

**1.** **基础理论**

观察和逻辑推理是科学的基础，数据自然界中观察方法与数据推理的理论和方法是主 要的研究内容，主要包括数据的存在性、数据测度、时间、数据代数、数据相似性、簇论 与数据分类等。

**2.** **实验方法与逻辑推理方法**

建立数据科学的实验方法，需要提出科学假说和建立理论体系，并通过这些实验方法 和理论体系开展数据自然界的探索研究，从而掌握数据的各种类型、状态、属性、变化形 式和变化规律，揭示自然界和人类行为现象之间的规律。

**3.** **领域数据学**

将数据科学的理论和方法广泛应用，开发出专门的理论、技术和方法，从而形成专门 领域的数据科学，例如脑数据学、行为数据学、生物数据学、气象数据学、金融数据学、 地理数据学等。





**大数据基础与应用**

**4.** **数据资源的开发方法和技术**

数据资源是重要的现代战略资源，具有巨大的价值，越来越凸显重要性，是继石油、 煤炭、矿产之后的最重要的资源之一。这是因为人类的社会、政治和经济都将依赖于数据 资源，而石油、煤炭、矿产等资源的勘探、开采、运输、加工、产品销售等无一不是依赖 数据资源的，离开了数据资源，这些工作都将无法开展。

**17.1.3** **数据科学的研究过程与体系框架**

**1.** **数据科学的研究过程**

● 从自然界中获得一个数据集。

● 对该数据集进行研究与探索，进而发现整体特性。

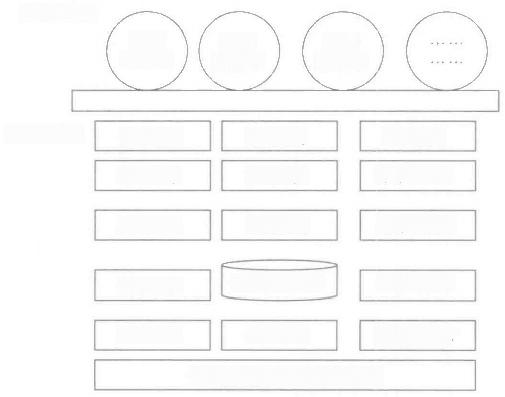
● 进行数据分析或数据实验。

● 发现数据规律。

● 将数据进行感知化等。

**2.** **数据科学的体系框架**

数据科学的体系框架如图17- 1所示。



数据学分类

宇宙 生命 行为

数据学 数据学 数据学

|  |  |  |
| --- | --- | --- |
| 数据学基础 数据获取 | 数据分析 | 数据感知 |
| 数据翻译 | 数据工程 | 计算可视化 |

数据分类 数据压缩 计算可视化

数据整合 数据森林全书 计算可视化

数据定义 数据分类 计算可视化

计算科学基本内容与基础理论

图17-1 数据科学的体系框架

数据科学的研究对象、研究目的和研究方法与计算科学、信息科学有本质的不同。数 据存于存储空间中，信息是自然界、人类社会及人类思维活动中存在和发生的现象，知识 是人们在实践中所获得的认识和经验。数据可以作为信息和知识的符号表示或载体，但数 据本身并不是信息或知识。数据科学的研究对象是数据，而不是信息，也不是知识。通过 研究数据来获取对自然、生命和行为的认识，进而获得信息与知识。

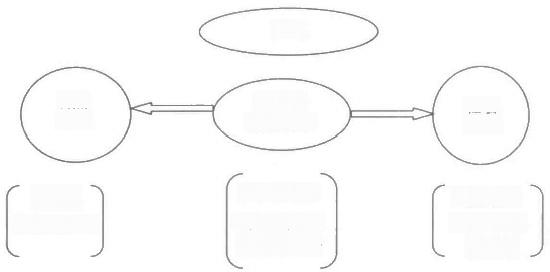
自然科学研究自然现象和规律，认识的对象是整个自然界物质的各种类型、状态、属





**第** **1** **7** **章** **数据科学与数据思维**

性及运动形式。行为科学是研究自然和社会环境中人的行为以及低级动物行为的科学，已 经确认的学科包括心理学、社会学、社会人类学和其他类似的学科。数据科学支持自然科 学和行为科学的研究。随着数据科学的进展，越来越多的科学研究将直接针对数据进行， 这将使人类通过认识数据来认识自然和行为。数据科学与其他学科的关系如图17-2所示。



人类

数据学 数据科学

数据自然界 存储在计算机 中任何数据

社会

科学

人类行为

社会、法律

和经济

自然 科学

自然界

宇宙和生命

图17-2 数据科学与其他学科的关系

人类探索自然界，用计算机处理人类的发现、人类的社会、自然与人的关系，在这个 过程中，产生了大数据，人类在不知不觉中创造了一个更复杂的数据自然界。人们生活在 现实自然界和数据自然界两个世界中，人、社会和宇宙的历史将变为数据的历史。人类可 以通过探索数据自然界来探索自然界，人类还需要探索数据自然界特有的现象和规律，这 是数据科学的任务。目前的所有的科学研究领域都可能形成相应的数据科学。

数据科学不仅将给科研和教学体制带来大幅度的变革，也会给科学与产业之间的关系、 科学与社会之间的关系带来大幅度的变革。信息时代，万物数化，许多学科已经和信息科 技深度融合，形成新的研究领域，如生物信息学、天体信息学、数字地球、计算社会学等。 一方面，用数据来研究科学已经是科学研究的主要手段之一；另一方面，大量的、非结构 化的数据，同样需要科学的手段、科学的研究数据。产业界在生产经营中积累丰富的数据， 学术界则有待于用实践检验模型和算法。数据科学为学术界和产业界的紧密衔接提供了纽 带和桥梁，促进了产、学、研深度融合。

**17.2** **大数据研究方式**

大数据的分析不能停留在仅仅获得概率分布结果，也不能满足于对细节问题的数据挖 掘，面对 PB 级以上的数据，需要更简单而有效的问题求解方法，要争取从大数据中获得 新知识与新价值。应优先选择预言性数据挖掘问题的研究， 一个重要的研究点是数据规模 大到何种程度，就可以解决以前解决不了的问题，进而实现社会科学变革式的巨大进步。

支持大数据研究与应用的是数据科学，数据科学是大数据研究的基础。目前颇受重视 的是大数据分析和大数据系统效率。工程上无法解决的问题就很自然地成为数据科学的研 究内容，大数据处理技术的进步又将促进数据科学的发展。



**大数据基础与应用**

大数据技术是在多种类的、大量的数据中快速获取信息的技术，将信息处理模式从以 计算为中心转向以数据为中心。大数据正以前所未有的速度颠覆人们探索世界的方法，大 数据的出现是对IT 领域的全面冲击。这是由于以往的模型与算法是基于小数据、小样本的 模型与算法，大数据带给我们的是一种全新的分析方式。

**17.2.1** **大数据分析的是全面的数据**

由于采集、存储和分析等数据技术工具限制，只能收集少量数据进行分析，为了使分 析更为简单，可将数据量缩减到最少。统计学的一个目的就是用尽可能少的数据来证实尽 可能多的发现，在大数据时代，有三个标志性的改变。第一个标志性的改变是，可以分析 更多的数据，有时候甚至可以处理与某个特别现象相关的所有数据，而不再依赖于随机采 样。第二个改变是，研究数据如此之多，以至于我们不再热衷于追求精确度。第三个转变 因前两个转变而促成，即我们不再热衷于寻找因果关系，取而代之的是寻找联系关系。

**1.** **样本分析法**

样本分析法采用了随机采样方法，目的是应用最少的数据来获得最多的信息。样本分 析法一直存在较大的问题。事实证明，问题的关键是选择样本时的随机性，样本选择的随 机性比样本数量更为重要。统计结果表明，随着采样随机性的增加，样本分析法的精确性 大幅提高，但却与样本数量的增加关系不大。

**2.** **随机采样**

随机采样开辟了一条收集信息的新途径。通过随机收集样本，可以利用较少的数据做 出高精准度的推断，即随机采样使用最少的数据来获得最多的信息。随机采样是在不可收 集和分析全部数据的情况下出现的选择方法，但是，它本身存在许多固有的缺陷。它的成 功完全依赖于采样的随机性，但是实现采样的绝对随机性非常困难， 一旦采样过程中偏离 绝对随机性，分析结果可能相差甚远。

(1)随机采样不适合考察子类别

随机采样不适合考察子类别的情况。因为一旦继续细分，随机采样结果的错误率将大 大增加。也就是说，当需要了解更深层次的细分领域的情况时，随机采样的方法就不可取。 在宏观领域起成功作用的方法在微观领域失去了作用。随机采样就像是模拟照片打印，远 看很清晰，但是一旦聚焦某个点，就会变得模糊不清。例如用采样的方法分析整个人口的 情况，正确率可达97%。对于某些事物来说，3%的错误率显然可以接受。但是无法得到一 些微观细节的信息，甚至还会失去对某些特定子类别进行进一步研究的能力。正态分布是 标准的。生活中真正有趣的事情经常藏匿在细节之中，但是，采样分析法却无法捕捉到这 些细节。

(2)需要严密的安排和执行

随机采样需要严密的安排和执行，只能从采样数据中得出事先设计好的问题的结果， 采样的数据不能回答突然意识到的问题。随机采样是一条捷径，随机采样方法并不适用于 一切情况，因为这种调查结果缺乏延展性，即调查得出的数据不可以重新分析以实现计划





**第17章** **数据科学与数据思维**

之外的目的。只研究样本而不是整体，有利有弊。能更快更容易地发现问题，但不能回答 事先未考虑到的问题。

**3.** **全数据模式**

由于处理能力和数据收集能力限制，导致了采样分析的应用。但现在计算机的处理能 力有了巨大的提升，应用的瓶颈不是计算能力而是数据采集能力，表现在采样的很多场景 是不可能收集到全部数据。

全数据模式是指样本等于总体，也可以称之为全样本。在信息处理能力受限的时代， 需要数据分析，但却缺少收集数据的工具，因此随机采样应运而生。现在，计算机可以轻 易地对这些数据进行处理。采样的目的就是用最少的数据得到最多的信息。当我们可以获 得大数据的时候，采样技术的研究也就无意义了。数据处理技术已经发生了巨大的变化， 采样方法需要跟上这种变化。

利用样本等于总体可以对数据进行深度研究，但是采样几乎无法达到这样的效果。例 如，初始使用的样本包含12000个数据，所以取得了不错的预测结果。随着收集的数据越 来越多，预测结果会越来越准确。

因为大数据技术是建立在掌握所有数据，至少是尽可能多的数据的基础上，所以可以 正确地考察细节并进行新的分析。在任何细微的层面，都可以用大数据去论证新的假设。 当然，毕竟在一个资源有限的时代，有些时候还可以使用样本分析法，但是更多时候，利 用掌握的所有数据成为最好的、可行的选择。

社会科学是被全数据模式撼动得最严重的学科。随着大数据分析取代了样本分析，社 会科学不再单纯依赖于分析经验数据。这门学科过去曾非常依赖样本分析、研究和调查问 卷。当记录下来的是人们的平常状态，也就不用担心在做研究和调查问卷时存在的偏见了。 现在，可以收集过去无法收集到的信息，不管是通过移动电话表现出的关系，还是通过 Twitter 信息表现出的感情，现在可以不再依赖抽样调查了。

**17.2.2** **重视数据的复杂性与弱化精确性**

在小数据时代，因为当时收集的数据很少，而不能使用更多的数据，所以需要越精确 越好，因此，追求精确度是合理的。但是，在大数据时代，数据量的限制正在逐渐消失， 而且通过无限接近样本等于总体的方式来处理数据，将获得极大的好处。如今快速获得一 个大概的轮廓和发展脉络，就要比严格的精确性数据更重要。

在使用抽样数据的时候，要容忍一定的误差。就算取得了全样本数据，也可能因为各 种原因而导致不精确，统计实践中对此已有相当多的案例。人类从未奢望通过数据分析取 得的多数结论是精确的，从来都要在信息混杂的情况下做出大概的决策，也就是说，需要 效率，不需要绝对精确。

**17.2.3** **关注数据的相关性而非因果关系**

除了纠结于数据的准确性、正确性、纯洁度和严格度之外，也应该容许一些不精确的





**大数据基础与应用**

存在。数据不可能完全对或完全错。当数据的规模以数量级增加时，这些混乱也就算不上 问题了。事实上，它甚至可以具有好处，因为当我们只想使用一小部分时，无须捕捉这么 多的知识细节。又因为可以用更快、更便宜的方式找到数据的相关性，并且效果往往更好， 而不必努力去寻找因果关系。当然在某些情况下，仍然需要精心策划的数据来做因果关系 研究和控制实验，如测试药物的副作用或设计关键的飞机部件。但是在日常情况下，知道 “是什么”就已经足够，不必非要弄清楚“为什么”。大数据的相关性比探讨因果关系更有 前景。

在哲学界，关于因果关系是否存在的争论已经持续了几个世纪。如果凡事皆有因果的 话，那么就没有决定任何事的自由了。如果所做的每一个决定或者每一个想法都是其他事 情的结果，而这个“其他事情”又是由其他原因导致的，以此循环往复，那么就不存在人 的自由意志这一说法了，即所有的生命轨迹都只是受因果关系的控制。

统计学的目标是从各种类型的数据中提取有价值的信息，给人以后见之明，但不强调 对事物的洞察力，不强调深度的知识。所以，从固有的统计思维突破到数据的思维是一重 大进步。例如，在大数据分析领域，更关心事物的相关性或者关联性。与传统的逻辑推理 研究注重条件和结果之间的因果关系不同，相关分析是为了找出数据集里隐藏的相互关系 网， 一般用支持度、可信度和兴趣度来衡量与表示。大数据分析的巨大驱动是商业盈利。 就像在著名的啤酒、尿布的例子中，以这种相关性来采取措施增加企业利润，而不用去深 究背后的内在规律和机制。

大数据时代的很多模型都是为了指导商业决策而设立，而商业决策通常会影响决策者 的利益。所以一个模型是否正确不是最重要的，重要的是决策者对这个模型有多大的把握， 决策者能否从这个模型中获利。所以大数据时代最为关键的应该是基于数据的模型能否说 服决策者据此进行决策，并且帮助决策者改善决策赚取相应的利润。前者表现为决策者愿 意将多少钱押在这个模型上，而后者表现为这个模型在现实中的表现如何。

因果性与相关性是人类思维和机器思维的区别，不能说机器数据分析更加关注因果性。

**17.3** **数** **据** **专** **家**

大数据给科学和教育事业的发展提供了前所未有的机会，同时也提出了前所未有的挑 战。随着大数据技术的发展与应用，需要大量数据科学家与数据处理工程师。

**17.3.1** **数据科学家**

数据科学家必须能够获取数据，并对其建模，以及懂得建立模型所需的数学理论。必 须能够发现并阐明数据揭示的问题，即从数据中发现合适的问题。数据科学家是指对数据 的数字化重现与认识，并在数据领域有一定贡献的人。这是最有吸引力的工作之一，利用 统计学获取并处理数据，从中得到有用信息并能图形化与可视化，并使人们得以理解，这 将是非常重要的功能。统计学家实际上是能够提取大数据的信息，然后展现给非数据专家





第17章 数据科学与数据思维

的那些人，又可以称之为数据科学家。

**1.** **数据科学家所具有的能力**

(1)统计分析能力

统计是数据科学的语法，至关重要。数据科学不只是获取数据，然后猜测其意义。它 也包含了假设检验，确保数据结论的有效性。统计现已成为一项基本技能，并没有被机器 学习或商务智能技术所取代。

(2)对数据的提取与综合能力

存储数据只是数据平台建设的一部分。在大数据计算中， MapReduce 方法基本上利用 非常庞大的计算集群进行分布计算。最流行的 MapReduce 开源实现方法就是 Hadoop 。 Hadoop 使敏捷数据分析成为可能。在软件开发中，敏捷表明更快的产品周期，开发商和消 费者之间更密切的互动。传统的数据分析需要极长的周转时间，可能要几个小时甚至几天 才能完成。但是在 Hadoop 帮助下可以很容易地建立集群，以执行大规模数据集的迅速计 算。更快的计算速度能更容易地测试不同的假设、不同的数据集和不同的算法。掌握大数 据处理的工具可以提升对数据的提取与综合能力。

(3)数据的可视化表示能力

可视化是数据的提炼和展现，是数据分析的初步探索工作，也是每个阶段的关键内容 之一。通过数据可视化技术可以发现规律与规则。 一张图片胜过千言万语，这就是数据可 视化的艺术。可视化分析是可视化工作与分析工作交替进行，获得满意的结果解释。

(4)知识基础

为了进行大数据的分析，数据科学家需要有数据库系统及数据管理的知识来应对大量 数据的导入和存储，同时必须掌握机器学习中的算法和模型处理预测性的需求。在整个过 程中，统计学的概念和人工智能的理论，都是选择正确、合适的分析方法和对分析结果进 行评估的重要依据。

**2.** **数据科学家的知识领城**

数据科学家应学习下列知识领域。

● 计算机科学：数据获取、数据解析、数据存放和数据安全。

● 数理统计学：数据分析、数据过滤、数据挖掘和数据优化。

● 图形设计学：显示数据结果，将数据表达成三维图形，以便更好地理解和利用。

● 人机交互学：在用户和数据之间建立有机联系，使得人对数据的使用更方便。

**3.** **数据科学家的特征**

(1)较强的探索精神

好奇心是数据科学家的必要素质，这与任何领域的科学家一样，数据科学家也需要具 备基本的好奇心和较强的探索精神，并能够将好奇心发展到极致。他们从数据中发现事物 的规律和原因。他们会从数据的不同角度来进行研究，进而发现别人所看不到的规律。

(2)较好的数学水平

数据科学家能够发现别人不易注意的东西。例如，刘某现在已经是一个大的媒体集团



**大数据基础与应用**

的首席数据科学家。在他第二天上班的时候，老板给了他一叠报告。他简单扫了一眼报告， 发现了一个关于回报率的计算错误。他花了一个小时验证了这个错误并算出正确答案。重 要的是，几百人都看过这份报告，肯定有一些相当好的分析师也看过这个报告，而只有他 看出了这个错误。在描述现实世界的时候，数学是一种非常有效的语言。优秀数据科学家 对这种语言具备一种天生的感觉。

(3)较强的技术能力

掌握最新的数据分析方法很重要，更重要的是要具备统计学的知识，并能够应用统计 学。简单地说，数据分析包括描述型分析和预测型分析。掌握预测型分析比描述型分析更 困难，这是因为预测型分析充满了不确定性。数据科学家能够驾驭不确定性，知道预测的 哪些部分是基于真实的观察，哪些因素是基于假设。他也知道，要使预测有效，哪些条件 必须满足，哪些因素会导致预测失效，哪些未知因素可能会导致预测错误等。数据科学家 能够量化风险，并设计一些小型实验来验证或者推翻某些假设。这不是一种数学技能，而 是在长期复杂的工作中，经过了无数次成功和失败之后总结和培养出来的一种技术能力。

(4)锲而不舍的精神

数据科学家在很多方面都能表现出持之以恒的特点。对于一个研究问题，看了一眼发 现了错误，花了一个钟头就验证出来了。实际上数据分析很少能这么快，这是由于不可能 轻易地揭示秘密，需要把秘密找出来。即便在最好的情况下，数据也常常不完整甚至存在 错误，而且大部分数据最后都和需要解决的问题无关。紧盯着这些噪声数据是一项单调无 趣的工作。除了上述内容之外，数据科学家还应该能够坚持表述自己的发现。

(5)经验和直觉

从什么地方开始着手了解数据，需要经验和直觉。数据科学家在数据技术的研究和实 践中积累了丰富的经验，对数据反应灵敏，并能够应用经验和直觉助推数据分析的完成。

(6)设计实验的场景

数据科学家应知道如何虚拟和假设场景，了解得到的数据。如果得不到数据，需要了 解其中的原因。如果从事数据科学的工作，就要求从专业背景上提出问题，并用数据和统 计方法进行检验，这样才能真正发现并构建知识。

**4.** **数据科学家的主要工作**

(1)数据处理平台的构建

数据处理平台包括公司的基础数据平台以及各个具体业务线的指标数据和日志数据平 台。这步工作的设计和技术选型紧密地依赖于后两步的分析需求。

(2)历史数据的分析挖掘

历史数据的分析挖掘包括各种产品线相关的业务分析、用户画像、用户行为分析、用 户留存分析等。类似的分析可以以图标或其他可视化的方式展现，其目的是使业务决策者 对于现状有清晰、系统和完整的认识，从而辅助其做出下一步的动作。

(3)数据驱动的预测性分析

数据驱动的预测性分析主要是建立推荐模型并且利用模型对未来的情况进行预测。在 计算广告中，点击率预估模型能被用来给特定的人和场景推荐合适的广告。这一部分的工





**第17章** **数据科学与数据思维**

作是大数据挖掘下最有意义的工作之一，也是与产品线联系最紧密的部分。

为了进行大数据的分析，首先面临知识范畴的挑战。数据科学家需要有数据库系统及 数据管理的知识来应对大量数据的导入和存储，同时必须掌握机器学习中的算法和模型处 理预测性的需求。在整个过程中，统计学的概念和人工智能的理论是选择正确、合适的分 析方法和对分析结果进行评估的重要依据。

**17.3.2** **数据处理工程师**

企业是开展大数据分析与应用的基本单位，是推动大数据应用的重要动力。大数据处 理工程师是在企业中的数据专家，是完成数据分析和数据工程的主要参加者和主持者， 应具备下述基本的能力。

● 整合企业数据的能力。

● 探索数据隐含价值的能力。

● 制订精确行动计划的能力。

● 进行精确、快速、实时行动的能力。

● 精确掌握数据处理技术。

**17.3.3** **大数据思维**

思维的产生发展与科学技术的发展、科学技术工具的产生发展密切相关，计算机的出 现催生了计算思维的产生与发展，而现在人们所关注的网络思维、系统思维，以及大数据 思维是计算思维的进一步发展与补充。

历来的商业变革都是由思维方式的转变开始的，旧的经济体制和传统的商业理念面临 新的商业思维逻辑的时候，如果大脑不能与时俱进，吸收并转变为顺应潮流的新思维，通 过新思维重新组织企业组织的战略、结构、文化和各种策略，那么貌似强大的体魄反而变 成了企业前进的累赘。

大数据思维是大数据时代的产物，是计算思维的最新的重要发展。大数据思维首先要 重视数据的全面性，而非随机的抽样性。其次是关注数据的复杂性，弱化精确性，要求对 一个大的框架、模糊的准确度趋势的判断。大数据是一种重新评价企业、商业模式的新方 法，数据成为核心的资产，并将深刻影响企业的业务模式，甚至重构其文化和组织。

对于数据科学来说，只考虑商业分析的数据支持，这还是小数据思维的思想。从金融、 运营商、政府等所完成的项目中可以发现，大数据已嵌入到整个行业中。利用新的、全面 的数据与新的证据不断修订假设是贝叶斯思维方式的基本理念，这也是大数据时代思维的 基本理念。

统计学方法论是进行大数据分析的重要基础。将统计学与大数据相融合将颠覆很多原 有的思维方式。在大数据时代，因为数据量很大，很可能找到新的关系，但是因为数据太 多，不一定能够理解为什么是这样。但有时可能找到内在的因果关系。在大数据时代，不 一定会知其所以然，但更方便大家知其然，原来找不到的相关关系现在找到了。





**大数据基础与应用**

**小结**

本章介绍了数据科学的定义和特征、数据科学的研究内容与研究过程、数据科学家和 数据工程师、大数据思维等内容。这些内容是本书的概括性的总结。通过上述内容学习， 可以了解数据科学的产生与发展，理解大数据思维的基本思想。



**参** **考** **文** **献**

[1]李国杰.大数据研究的科学价值[J]. 中国计算机学会通讯，2012,8(9):8-15. [2]陈国良.计算思维导论[M]. 北京：高等教育出版社，2012.

[3]赵致琢.计算科学导论[M]. 北京：科学出版社，2004.

[4]Tony Hey. 第四范式：数据密集型科学发现[M]. 潘教峰，张晓林，等译.北京： 科学出版社，2012.

[5]董荣胜，古天龙.计算机科学与技术方法论[M]. 北京：人民邮电出版社，2002.

[6]马帅，李建欣，胡春明.大数据科学与工程的挑战[J]. 中国计算机学会通讯，2012, 8(9):22-28.

[7]俞宏峰.大规模科学可视化[J]. 中国计算机学会通讯，2012,8(9):29-36.

[8]黄伯仲.超大规模数据可视化分析十大挑战[J]. 中国计算机学会通讯，2012,8(9): 38-43.

[9]周涛.个性化推荐的十大挑战[J]. 中国计算机学会通讯，2012,8(7):48-57.

[10]李航.从大数据中挖掘什么[J]. 中国计算机学会通讯，2013,9(6):36-39.

[11]陈明.大数据问题[J]. 计算机教育，2013(5):103-110.

[12]陈明.数据密集型科研第四范式[J]. 计算机教育，2013(9):103-110. [13]陈明.NoSQL 数据库系统[J]. 计算机教育，2013(11):107-111.

[14]陈明.分布式系统设计的CAP 理论[J]. 计算机教育，2013(15):109-112. [15]陈明.MapReduce 分布编程模型[J]. 计算机教育，2014(1):104-117.

[16]陈明.大数据分析[J]. 计算机教育，2014(5):122-126.

[17]陈明.分布计算应用模型[M]. 北京：科学出版社，2009.

[18]艾伯特-拉斯洛·巴拉巴西.爆发——大数据时代预见未来的新思维[M]. 北京： 中国人民大学出版社，2012.

[19]Martin Klubeck. 量化——大数据时代的企业管理[M]. 吴海星，译.北京：人民 邮电出版社，2013.

[20]张鑫.Hadoop 源代码分析[M]. 北京：中国铁道出版社，2013.

[21]邹贵金.MongoDB管理与开发实战详解[M]. 北京：中国铁道出版社，2013.

[22]万川梅，谢正兰.Hadoop应用开发实战详解[M]. 北京：中国铁道出版社，2013. [23]Tom Wbite.Hadoop 权威指南[M]. 周敏奇，译.北京：清华大学出版社，2011.

[24]Keith Willetts. 数字经济大趋势[M]. 徐俊杰，裴文斌，译.北京：人民邮电出版 社，2013.

[25]阳振坤，张清，王勇，等.大数据的魔力[J]. 中国计算机学会通讯，2012,8(6): 17-21.



**大数据基础与应用**

[26]周晓方，陆嘉桓，李翠平，等.从数据管理视角看大数据挑战[J]. 中国计算机学 会通讯，2012,8(9):16-20.

[27]陈明.大数据概论[M]. 北京：科学出版社，2014.

[28] 王文生，陈明.大数据与农业应用[M]. 北京：科学出版社，2011.

[29]Nancy Lynch,Seth Gilbert.Brewer's conjecture and the feasibility of consistent, available,partition-tolerant web services[J].ACM SIGACT News,2002,33(2):51-59.

[30]Brewer E.Towards Robust Distributed Systems.19th Acm Symposium on Principles of Distributed Computing,2000.

[31]吴锐.有容乃大——大规模数据云端存储[J]. 中国计算机学会通讯，2012,8(6): 26-28.

[32]陆嘉桓.Hadoop 实战[M]. 北京：机械工业出版社，2012.

[33]佐佐木达也.NoSQL 数据库入门[M]. 罗勇，译.北京：人民邮电出版社，2012.

[34] 毛国君，段立娟.数据挖掘原理与算法 [M]. 第3版.北京：清华大学出版社， 2016.

[35]王星，等.大数据分析：方法与应用[M]. 北京：清华大学出版社，2013.

[36]鲍亮，李倩.实战大数据[M]. 北京：清华大学出版社，2014.

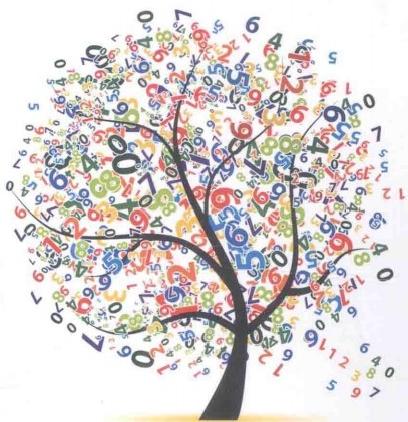
[37]黄宜华.深入理解大数据[M]. 北京：机械工业出版社，2014.

[38]深圳国泰安教育技术股份有限公司大数据事业部群，中科院深圳先进技术研究 院——国泰安金融大数据研究中心.大数据导论：关键技术与行业应用最佳实践[M]. 北京： 清华大学出版社，2015.

[39]Anand Rajarman,Jaffrey David Ullman. 大数据：互联网大规模数据挖掘与分布式 处理[M]. 王斌，译.北京：人民邮电出版社，2012.

[40]孙大为，张广艳，郑伟民.大数据流式计算：关键技术与系统实例[J].软件学报， 2014(4):839-862.





**大数据**

**基础与应用**

大数据正在改变我们的一切。

大数据的研究已经成为计算机科学与技术研究的 新热点，并显示出日益强大的生命力和吸引力。科学 大数据催生了数据密集型知识发现的第四科学研究范

式的出现。

大数据是继20世纪末、21世纪初互联网蓬勃发

展以来的又一轮新的IT工业革命。



微信订阅号：kjjbbooks

微信服务号：kjbbooks-s