

数据仓库与数据挖掘

Data Warehousing and Data Mining

武森 高学东 [德] M.巴斯蒂安 著



冶金工业出版社

<http://www.cnmp.com>

数据仓库与数据挖掘

武 森 高学东 著
[德]M.巴斯蒂安

北 京

冶 金 工 业 出 版 社

2003

图书在版编目(CIP)数据

数据仓库与数据挖掘/武森等著. —北京:冶金工业出版社,2003.9

ISBN 7-5024-3294-9

I. 数… II. 武… III. 数据库系统 IV. TP311.13

中国版本图书馆 CIP 数据核字(2003)第 037843 号

出版人 曹胜利(北京沙滩嵩祝院北巷 39 号,邮编 100009)

责任编辑 张 卫(010-64027930,E-mail:bull2820@sina.com) 郭庚辰

美术编辑 李 心 责任校对 刘 倩 责任印制 牛晓波

北京市铁成印刷厂印刷;冶金工业出版社发行;各地新华书店经销

2003 年 9 月第 1 版,2003 年 9 月第 1 次印刷

850mm×1168mm 1/32;12.375 印张;329 千字;379 页;1-4000 册

30.00 元

冶金工业出版社发行部 电话:(010)64044283 传真:(010)64027893

冶金书店 地址:北京东四西大街 46 号(100711) 电话:(010)65289081

(本社图书如有印装质量问题,本社发行部负责退换)

前 言

随着市场竞争的日趋激烈,信息对于企业生存和发展所起的作用越来越重要。由于计算机技术的普遍应用,承载信息的数据随着时间推移而不断积累并与日俱增,使得企业难以从巨量数据中找到真正有用的决策信息,因此企业迫切需要新的技术和方法从巨量数据中提取有价值的信息或知识。数据仓库与数据挖掘是20世纪90年代发展起来的新技术。数据仓库是一种针对大数据集进行数据组织与管理的技术,专门用于支持分析型的数据查询,而数据挖掘则是从大量数据中寻找蕴涵其中但尚未发现的知识。

本书是作者在多年研究与教学工作的基础上完成的,书中全面系统地介绍了数据仓库和数据挖掘领域相关内容的主要思想、基础理论、核心技术与方法。全书共分14章,其中1~8章介绍了数据仓库的基本概念、建模方法、体系结构、主题数据组织与数据预处理技术、OLAP及数据仓库的规划与管理;9~14

前 言

章深入讨论了数据挖掘的基本概念、数据准备、数据挖掘技术(包括聚类分析、分类发现和关联规则)及数据挖掘的发展与应用。

本书适用于数据仓库与数据挖掘领域的研究和应用人员,也可作为相关专业研究生和高年级本科生的教学用书或教学参考书。

在本书的写作过程中,王莹参加了第7、8章的写作,龙誉、王文贤分别参加了第12章、第13章的写作,第14章由宫雨写作完成,在此表示感谢!崔焕荣在书稿公式编辑方面做了大量的工作,在此一并表示感谢!在本书的写作过程中,作者参阅了大量文献,在此向文献作者表示感谢!

由于作者水平所限,加之数据仓库与数据挖掘是一个新兴的研究领域,有些专业术语尚未达成共识,书中叙述若有不妥之处,诚恳希望同行和读者批评指正,以便今后改正和完善。

作 者

2002年12月

目 录

1 数据仓库概述	1
1.1 数据仓库的产生.....	1
1.2 数据仓库的含义.....	3
1.3 数据仓库的特征.....	5
1.4 操作型数据库系统与数据仓库.....	7
1.5 数据仓库的基本结构.....	9
本章要点	11
2 数据仓库相关概念	12
2.1 主题.....	12
2.2 粒度.....	13
2.3 维度.....	15
2.4 数据立方体.....	17
2.5 联机分析处理.....	20
2.6 数据集市.....	23
本章要点	25
3 多维数据模型	27
3.1 实体-关系模型与多维模型.....	27
3.1.1 实体-关系模型	27
3.1.2 多维数据模型	28
3.1.3 多维数据模型的优势	30
3.2 星形模式.....	32
3.2.1 星形模式的含义	32

3.2.2	主码、外码和代理码	34
3.2.3	事实表	36
3.2.4	维表	38
3.3	星形模式的种类	39
3.3.1	简单星形模式	39
3.3.2	星系模式	39
3.3.3	星座模式	41
3.3.4	二级维表	42
3.3.5	雪花模式	42
3.4	数据仓库的总线型结构	44
3.4.1	总线型结构的含义	44
3.4.2	统一的维	45
3.4.3	统一的事实	46
3.4.4	数据仓库的总线	47
	本章要点	49
4	数据仓库的体系结构	51
4.1	体系结构的内容	51
4.1.1	总体框架	51
4.1.2	技术体系结构	52
4.2	相关的数据存储	54
4.2.1	数据源	55
4.2.2	主题数据	56
4.2.3	预处理数据	58
4.2.4	查询服务数据	60
4.3	相关的数据服务	61
4.3.1	后台数据预处理	61
4.3.2	前台数据查询服务	62
4.4	相关的数据管理——元数据	64
4.4.1	元数据的含义	64
4.4.2	元数据的内容	65

4.4.3 元数据的工作流程	67
本章要点	69
5 数据仓库的数据组织	71
5.1 事实表和维表的设计	71
5.1.1 事实数据和维数据的区分	71
5.1.2 事实表的设计	72
5.1.3 维表的设计	75
5.1.4 常见维设计举例	75
5.2 数据聚集的设计	79
5.2.1 数据聚集的含义	79
5.2.2 数据聚集的创建方法	80
5.3 数据仓库中的索引	82
5.3.1 传统的数据库索引技术	82
5.3.2 事实表的索引	84
5.3.3 维表的索引	85
5.3.4 数据仓库索引举例	86
5.3.5 数据仓库索引新技术	88
5.4 数据库的物理设计	91
5.4.1 物理设计的内容	91
5.4.2 数据库对象的命名规范	92
5.4.3 物理模型的建立	93
5.4.4 数据仓库的数据量估计	95
本章要点	96
6 数据仓库的数据预处理	98
6.1 数据的净化	98
6.1.1 数据质量与数据净化	98
6.1.2 数据净化的方法	100
6.1.3 其他提高数据质量的方法	102
6.2 数据预处理计划	104
6.2.1 初步计划	104

6.2.2 详细计划	106
6.3 维表的数据预处理	108
6.3.1 维表的基本数据预处理	108
6.3.2 代理码的分配	110
6.3.3 维表的变更处理	111
6.4 事实表的数据预处理	113
6.4.1 事实表的基本数据预处理	113
6.4.2 非代理码的替换	115
6.4.3 数据聚集的更新维护	116
本章要点	117
7 联机分析处理——OLAP	119
7.1 基于多维模型的数据分析	119
7.1.1 OLAP 的含义与特征	119
7.1.2 OLAP 的基本操作	120
7.2 数据仓库建设与 OLAP 应用例	123
7.2.1 订货分析主题	123
7.2.2 数据源分析	124
7.2.3 订货分析的星形模式	125
7.2.4 订货分析的 DTS 包	130
7.2.5 订货分析的 OLAP 实践	135
本章要点	138
8 数据仓库的规划与管理	139
8.1 数据仓库系统的生命周期	139
8.2 项目的准备	141
8.3 项目团队的组织	142
8.4 项目的进度安排	144
8.5 项目的文档管理	145
本章要点	146
9 数据挖掘概述	148
9.1 KDD 与数据挖掘	148

9.2	KDD 过程	150
9.3	数据挖掘的任务	152
9.4	数据仓库与数据挖掘	153
	本章要点	155
10	数据挖掘的数据准备	157
10.1	数据准备概述	157
10.1.1	数据准备的内容	157
10.1.2	与数据仓库的比较	159
10.2	数据的应用变换	161
10.2.1	异常值处理	161
10.2.2	数据标准化	164
10.2.3	数据泛化	166
10.2.4	数据聚集	169
10.3	数据的精简	170
10.3.1	属性子集选择	170
10.3.2	主成分分析	171
10.3.3	离散小波转换	172
10.3.4	回归方法	174
10.3.5	数据抽样	175
	本章要点	177
11	聚类分析	179
11.1	聚类分析概述	179
11.1.1	聚类分析的含义	179
11.1.2	聚类方法的分类	180
11.1.3	数据挖掘应用对聚类分析的要求	181
11.2	差异度的计算方法	183
11.2.1	区间变量	183
11.2.2	二态变量	184
11.2.3	分类变量	186
11.2.4	序数变量	187

11.2.5 计算中的其他问题	188
11.3 分割聚类方法	190
11.3.1 分割聚类方法概述	190
11.3.2 k-means 算法	191
11.3.3 PAM 算法	194
11.3.4 CLARA 算法	196
11.3.5 CLARANS 算法	197
11.4 层次聚类方法	200
11.4.1 层次聚类方法概述	200
11.4.2 层次聚类基本算法	202
11.4.3 BIRCH 算法	204
11.4.4 CURE 算法	208
11.5 基于密度的聚类方法	211
11.5.1 基于密度的聚类方法概述	211
11.5.2 DBSCAN 算法	213
11.5.3 OPTICS 算法	215
11.5.4 基于密度和网格的 CLIQUE 算法	218
11.6 高维稀疏聚类 CABOSFV 算法	221
11.6.1 算法的主要思想	221
11.6.2 算法的概念基础	223
11.6.3 算法的聚类过程	225
11.6.4 算法举例	227
本章要点	231
12 分类发现	233
12.1 分类发现概述	233
12.1.1 分类发现的含义与过程	233
12.1.2 分类模型的评估标准	234
12.1.3 分类发现的主要方法	236
12.2 决策树算法	237
12.2.1 决策树算法概述	237

12.2.2	决策树的生成	238
12.2.3	决策树生成举例	240
12.2.4	决策树剪枝举例	243
12.2.5	从决策树中提取规则	244
12.3	ID3 算法	245
12.3.1	ID3 算法的决策属性选择方法	245
12.3.2	ID3 算法示例	246
12.3.3	树的剪枝	249
12.3.4	ID3 的几种改进方法	250
12.4	ID3 改进算法 CAMM	251
12.4.1	CAMM 算法的概念基础	252
12.4.2	CAMM 算法的处理过程	254
12.4.3	CAMM 算法决策树生成举例	255
12.4.4	CAMM 算法的规则提取	259
12.4.5	决策树算法与数据仓库技术的集成	261
12.5	贝叶斯分类	264
12.5.1	贝叶斯原理	265
12.5.2	简单贝叶斯分类	266
12.5.3	贝叶斯信念网络	269
12.6	基于神经网络 BP 算法的分类	271
12.6.1	多层前馈神经网络	271
12.6.2	BP 算法	273
12.6.3	BP 算法的解释	275
12.6.4	其他分类算法	276
	本章要点	278
13	关联规则发现	281
13.1	关联规则概述	281
13.1.1	关联规则的基本概念	281
13.1.2	描述关联规则的参数	282
13.1.3	关联规则分类	285

13.1.4 发现关联规则的过程	286
13.2 Apriori 算法	286
13.2.1 Apriori 算法概述	287
13.2.2 Apriori 性质与算法步骤	288
13.2.3 Apriori 算法举例	290
13.2.4 由频繁集产生关联规则	293
13.2.5 Apriori 算法的几种优化方法	294
13.3 FP-growth 方法	297
13.3.1 FP-growth 方法的概念与步骤	297
13.3.2 FP-tree 的建立	298
13.3.3 在 FP-tree 上挖掘关联规则	299
13.4 多层与多维关联规则	302
13.4.1 概念树	302
13.4.2 自上而下挖掘多层关联规则	303
13.4.3 自下而上挖掘多层关联规则	307
13.4.4 多维关联规则	310
13.4.5 数值属性的离散化	311
13.5 具有利润约束的关联规则	314
13.5.1 利润约束	315
13.5.2 具有利润约束的频繁集	316
13.5.3 具有利润约束的关联规则发现算法	318
13.5.4 算法示例	320
13.5.5 其他约束条件	322
本章要点	325
14 数据挖掘的发展与应用	327
14.1 分布式数据挖掘	327
14.1.1 分布式数据挖掘简介	327
14.1.2 分布式数据挖掘系统	329
14.1.3 研究现状	331
14.2 分布式数据挖掘算法	332

14.2.1 分布式关联规则	333
14.2.2 分布式分类算法	337
14.3 数据挖掘软件发展	340
14.3.1 系统功能的发展	341
14.3.2 应用模式的发展	342
14.4 数据挖掘标准	344
14.4.1 过程标准	344
14.4.2 实现标准	351
本章要点	358
名词索引	360
参考文献	372

1 数据仓库概述

随着市场竞争的日趋激烈,信息对于企业的生存和发展发挥着越来越重要的作用。由于计算机技术的普遍应用,承载信息的数据随着时间的推移而不断增长,并且分布在不同的系统平台上,具有多种存储形式。能否从纷繁复杂、大量沉淀的数据环境中得到有用的决策信息,及时做出正确的分析与决策,已成为企业生存与发展至关重要的环节。自从20世纪70年代提出决策支持的概念以来,人们在决策支持系统(Decision Support System, DSS)理论及应用上做了大量的研究工作,并且在企业决策中发挥了积极的作用。但现有的大部分决策支持系统是基于传统数据库基础之上的,随着企业数据量的不断增加,需要对原有的信息进行提炼和加工,需要为企业领导提供集成化和历史化的数据,需要为企业全局的战略决策和长期趋势分析提供更有效的支持。然而,传统的数据库管理系统因自身的局限性已无法满足决策支持系统对数据的要求。因此,一种适用于决策支持系统的数据组织与管理技术——数据仓库技术(Data Warehouse)应运而生,并逐渐成为支持分析与决策的重要技术。

1.1 数据仓库的产生

20世纪70年代, Micheal. Scott. Morton 提出了决策支持的概念,人们力图用现有的数据进行深层次的分析和推理,为决策者提供决策所需的信息。于是,在管理信息系统的基础上,以数据分析和建模定量分析为基础的决策支持系统发展起来了。传统的数据库技术作为数据管理手段,主要用于联机事务处理(On-Line Transaction Process, OLTP),在这样的数据库中保存的是大量的日常业务数据。它在数据共享、数据与应用程序的独立性、维护数

据的一致性和完整性及数据的安全保密性等方面提供了有效的手段。当它与分析型应用结合时,却出现了许多问题。

首先,决策支持系统为掌握充分的信息,需要访问大量的企业内部数据和外部数据。在一个企业的各部门中,可能存在不同的操作型管理信息系统,如销售管理系统、财务管理系统、物资管理系统等。这些操作型应用系统也可能依赖于不同的数据管理平台,常见的有 Oracle、Sybase、SQL Server、Foxpro、Excel 等。在传统系统中,支持决策的内部数据就来源于这些不同的数据源,它们在数据的定义及组织方式上都可能不同。因此,决策者为获取相应的数据就必须熟悉不同的系统环境和数据定义,花费大量的时间理解数据的具体含义和相互间的关系。这样基于不同的数据源支持决策,不仅对决策者本身的素质要求非常高,而且信息的提取也是非常不方便的。

其次,传统数据库中的大量数据是事务型数据,即该数据是对每一项工作、管理对象的具体的、细节性的描述。在销售业务系统中的数据可能表现为一份合同、一张收款单、一张发票或一张货票等。而企业决策层人员所关心的不可能是这类细节性信息,决策支持系统需要的是综合的、总结性的数据,其特点是涉及的数据量很大,甚至可能涉及到上千条、上万条发票或其他记录,但往往并不需要细节数据或个体数据。决策支持系统的目的在于为决策者提取有用的信息,而决策者不可能也不应该花费时间浏览所有细节数据或个体数据。因此,基于传统的操作型数据库不适合建立分析型应用系统。

再次,事务处理型应用和分析决策型应用对数据库系统的性能要求不同。事务处理型应用的特点是数据存取操作频率高,每日进行成千上万次的输入、修改等记录操作,但每次操作处理的时间短,一般是多个用户分时共同使用系统资源,采用传统数据库系统事务处理型应用运行良好。而在分析决策型应用中,为了获取综合性的、有用的决策信息,应用系统可能需要连续运行几个小时、甚至更多的时间进行必要的计算,大量地占用系统资源。如果

将分析决策型应用与事务处理型应用共同放在同一数据库系统环境中,必然引起系统资源紧张,甚至使事务处理型应用瘫痪。

最后一个问题是在传统数据库中保存和管理的一般是当前数据,即使部分历史数据通过备份或历史数据库等形式保存了下来,却普遍被束之高阁,并没有得到充分的利用。而决策支持系统不仅需要当前的数据信息,而且还要求有大量的历史数据,尤其是对历史数据进行分析和比较,找出企业发展变化的趋势。就这一点而言,传统数据库系统也不能满足分析决策型应用的需要。

基于以上的论述可以得出结论,在事务处理型应用环境中直接构建分析决策型应用是不可行的。在这种情况下,面向决策分析型应用而组织和存储数据的数据仓库技术便应运而生。随着市场竞争的日趋激烈,企业经营行为逐渐发生改变,对分析决策的要求越来越高。不论是从效率而言,还是从有效性而言,建立在事务处理系统应用环境中的分析决策系统都无法满足现代企业的需要。为了提高分析决策的效率和有效性,面向分析决策型应用的数据处理及其数据必须与事务处理型应用处理及其数据分离,即必须把分析决策型数据从事务处理系统应用环境中分离开来,建立单独的分析决策型应用环境。数据仓库正是为了构建这种新的分析决策型应用环境而出现的一种数据存储和组织技术,并成为研究数据管理技术的新课题。

1.2 数据仓库的含义

我们现在称之为“数据仓库”的这一技术,最早发轫于20世纪80年代初W. H. Inmon的研究,并存在于其“记录系统”、“本原数据”(Atomic Data)、“决策支持数据库”等研究专题中^[1]。数据仓库的概念是W. H. Inmon在其《建立数据仓库》一书中提出的,目前它被认为是解决信息技术(IT)在发展中一方面拥有大量数据,另一方面有用信息却很贫乏这种不正常现象的综合解决方案。W. H. Inmon曾对数据仓库做了这样的描述:“数据仓库是90年